REPOR[illegible]

1 AGENCY USE ONLY (Leave blank) | 2. REPO[illegible] | T TYPE AND DATES COVERED ANNUAL 01 Nov 92 TO 31 Oct 93

(2)

4 TITLE AND SUBTITLE

PSYCHOPHYSICS OF COMPLEX AUDITORY AND SPEECH STIMULI

5 FUNDING NUMBERS

F49620-93-1-0033

61102F
2313
AS

6 AUTHOR(S)

Dr Richard E. Pastore

7 PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Dept of Psychology
State University of New York
P.O. Box 6000
Binghamton, NY 13902-6000

8 PERFORMING ORGANIZATION REPORT NUMBER

AFOSR-TR- 94 0108

9 SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

AFOSR/NL
110 Duncan Avenue, Suite B115
Bolling AFB DC 20332-0001

Dr John F. Tangeny

10 SPONSORING/MONITORING AGENCY REPORT NUMBER

11 SUPPLEMENTARY NOTES

12a DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release;
distribution unlimited

12b DISTRIBUTION CODE

13 ABSTRACT (Maximum 200 words)

A major focus on the primary project is to use of different procedures to provide converging evidence on the nature of perceptual spaces for speech categories. Completed research examined initial voiced consonants, with results providing strong evidence that different stimulus properties may cue a phoneme category in different vowel contexts. Thus, /b/ is cued by a rising second formant (F2) with the vowel /a/, requires both F2 and F3 to be rising with /i/, and is independent o the release burst for these vowels. Furthermore, cues for phonetic contrasts are not necessarily symmetric, and the strong dependence of prior speech research on classification procedures may have led to errors. Thus, the opposite (falling F2 & F3) transitions lead somewhat ambiguous percepts (i.e., not /b/) which may be labeled consistently (as /d/ or /g/), but requires a release burst to achieve high category quality and similarity to category exemplars). Ongoing research is examining cues in other vowel contexts, and isusing additional procedures to evaluate the nature of interaction between cuse for categories of both speech and music.

14 SUBJECT TERMS

15 NUMBER OF PAGES

16 PRICE CODE

17 SECURITY CLASSIFICATION OF REPORT	18 SECURITY CLASSIFICATION OF THIS PAGE	19 SECURITY CLASSIFICATION OF ABSTRACT	20 LIMITATION OF ABSTRACT
(U)	(U)	(U)	(UL)

Psychoacoustics & Auditory Cognition Laboratory

Department of Psychology
SUNY - Binghamton
Binghamton, NY 13902

1993 Annual Progress Report

AFOSR Grants F49609310033

Richard E. Pastore, Project Director

1993 Annual Report for AFOSR Grants F49609310033 and F49609310327.

Organization of this Report

This report is intended to provide a sampling or snap-shot of the status of research program supported by the Air Force Office of Scientific Research under Grant F49609310033 and supplemental AASERT award F49609310327. Instead of preparing a special report describing each of the facets of the ongoing and completed research, this report is a compilation of manuscripts describing the various major facets of the research. These documents have a general organization based upon status in relation to publication in peer-reviewed journals.

The first section of this report contains one manuscript which currently is under review.

The second section contains four completed manuscripts which are about to be submitted for review. These manuscripts range in completion from being essentially ready to be dropped in the mail (e.g., Cho et al.) to requiring a little more critical reading and fine tuning before submission (e.g., Hall et al.).

The third section contains detailed reports on two on-going projects which are not described in any of the other four sections of this report. These reports provide a detailed introduction to the study, summarize the results obtained to date and, provide a discussion of those results. These reports are intended to be the basis of subsequent manuscripts describing the results.

The final section of this report contains papers or posters presented at professional meetings (Acoustical Society of America and Psychonomic Society), where the research topic is not otherwise covered in any of the other sections of this report. For example, Feature Integration Theory and Illusory Conjunctions are the focus of completed and ongoing research. However, since a manuscript or detailed report is at least a month away from reaching a stage adequate for public access, the poster has been included. As a contrary example, the contents of the Acoustical Society papers by Hall and by Cho (listed below) are covered thoroughly in the completed manuscripts which these individuals as first authors. Therefore, no added effort was made to provide a transcription of these oral presentation.

The manuscripts, reports, papers, and posters (cited above) describe primarily the current findings, but most also represent the basic subject and approach to continuing research.

Patent Statement

The research completed to date has not resulted in any findings or developments which are appropriate for any type of patent application, and no such application has been sought.

Research Overview

The major objectives of the ongoing research projects involved the delineation of the nature of the processes which determine the perception of complex acoustic stimuli. The research has not developed any new techniques for investigating perception, but is unique in utilizing a set of established procedures to provide a comprehensive picture of perception and converging evidence for the nature and role of specific cues in perception. The selection of stimulus class (e.g., speech, music, tones) for a given set of experiments is based upon being able to most effectively address a given critical research question.

The manuscripts and specific reports contained in this document attest to the success and the importance of not only the completed research, but also the on-going projects. These statements have been prepared in the forms appropriate for standard scientific peer-review. If needed, this document, or the next annual document, can be modified to include the applied implications of the research findings.

Research Bibliography

Manuscripts Under Review

[1]Li, X-F., & Pastore, R.E. Perceptual Constancy of a Global Spectral Property: Spectral Slope Discrimination. Journal of the Acoustical Society of America, (accepted pending revisions).

Manuscripts about to be submitted

[1]Acker, B.E., Pastore, R.E., & Hall, M.D., Within-category discrimination of musical chords: Perceptual magnet or anchor? Perception & Psychophysics.

[1] Cho, J.L., Hall, M.D., & Pastore, R.E. Normalization of musical instrument timbre. Journal of Experimental Psychology: Human Perception & Performance.

[1] Hall, M.D., & Pastore, R.E. Effects of stimulus complexity on the perceptual organization of musical tones. Journal of Experimental Psychology: Human Perception & Performance.

[1] Huang, W., Hall, M.D., & Pastore, R.E. Mapping percepts in the major variant of the octave illusion. Perception & Psychophysics.

Paper Presentations

[1,3]Hall, M.D., & Pastore, R.E. (1993) An Auditory Analogue to Feature Integration. Psychonomic Society, Washington, D.C. Nov. 4, 1993. [Poster Presentation]

[1] Pastore, R.E. (1993) Implicit assumptions in modeling higher level auditory processes. Journal of the Acoustical Society of America, 93, 2307. [Abstract of Invited paper]

[2] Huang, W., Hall, M.D., & Pastore, R.E. (1993) An illusion based on dichotic fusion of harmonically related tones. Journal of the Acoustical Society of America, 93, 2316. [Abstract of Poster]

[2] Cho, J.E., Hall, M.D., & Pastore, R.E. (1993) Stimulus properties critical to normalization of instrument timbre. Journal of the Acoustical Society of America, 93, 2402. [Abstract of Paper]

[2] Li, X-F, & Cho, J. (1993) An exploration of phoneme structure and models of classification for place of articulation. Journal of the Acoustical Society of America, 93, 2390. [Abstract of Paper]

Detailed Reports on Work in Progress (when not summarized above)

Pastore, R.E., Farrington, S., & Jassal, S. Measuring the DL for identification of order of onset for complex auditory stimuli

Pastore, R.E., Acker, B.A., Cho, J., Li, X-F., & Farrington, S. Exploration of the perceptual structure of cues for place of articulation

---

[1] Manuscript included in this report.

[2] Paper presentation NOT included in this report. More complete description of work included in this report

[3] Major aspects of this research is continuing

Research Staff

Faculty

Richard E. Pastore, Ph.D.

Graduate Students

Xiaofeng (Sheldon) Li, Ph.D.[4]
Michael Hall, MA
Wenyi Huang, MA
Jennifer Cho, MA[5]
Barbara Acker, BA[6]
Shannon Farrington, BA
Sajni Jassal[7]

Undergraduate Students

Denise Rotavera
Laura Peyser
Ellen Hoffman
Shannon Farrington[8]

[4] Sheldon Li received his Ph.D. from our program, but has continued to work on projects in laboratory. Dr. Li recently completed post-doctoral work with Dr. Charles Watson at Indiana University. He now is with the Department of Electrical and Computer Engineering, The Johns Hopkins University

[5] Jennifer Cho currently is a part-time student while working as a co-op (intern) on a Navy project with IBM-Owego (Loral-Owego)

[6] Barbara Acker is supported by AASLP/F award. She will complete her MA in 1994.

[8] Shannon Farrington worked on the project both as a SUNY-Cortland undergraduate and as a Binghamton graduate student

[7] Sajni Jassal, a graduate student in another laboratory, worked on this project over the Summer of 1993

Perceptual Constancy of a Global Spectral Property: Spectral Slope Discrimination

Xiaofeng Li* and Richard E. Pastore

Department of Psychology
and
Center for Cognitive and Psycholinguistic Sciences
State University of New York at Binghamton

Running head: SPECTRAL SLOPE DISCRIMINATION

* Currently affiliated with the Center for Speech Processing, Department of Electrical and Computer Engineering, The Johns Hopkins University.

Abstract

The current study investigated the perceptual constancy of spectral slope discrimination when the fundamental frequency and spectral shape of the stimuli were varied across to be discriminated stimuli on a single trial. The three stimulus variables, all of which were global or emergent properties of a complex sound, represented two sound source properties and a filter property. A stimulus was synthesized by passing a source spectrum through a filter transfer function according to the source-filter model of complex sound production. Four experiments were conducted in this study. Experiment 1 examined the effect of the difference in overall stimulus level on spectral slope discrimination. Experiments 2 and 3 investigated, respectively, the effects of variations in the fundamental frequency and a filter property on spectral slope discrimination. Experiment 4 was designed to resolve two issues raised in the preceding experiments. The current study showed a significant performance decrement in spectral slope discrimination when a second source property--fundamental frequency--was varied. However, little detrimental effect was observed when the filter property--spectral shape--was varied. The study supported claims that listeners treat source properties as a unit which is relatively independent of filter properties.

PACS numbers: 43.66.Jh, 43.66.Lj

Journal of the Acoustical Society of America

[accepted pending revsion]

One goal of psychoacoustics research is to evaluate the important logical possibility that principles discovered and results obtained from studies using simple stimuli (e.g., pure tones and noise bursts) can be extended to explain speech perception. Unfortunately, evidence has been accumulated over years indicating that, other than some very general findings (e.g., simultaneous masking), such efforts typically are not very successful (Pastore, 1981; Watson, 1991; Watson, Qiu, Chamberlain, & Li, 1993). For example, Christopherson and Humes (1992) examined the relationship between listeners' abilities to process a wide variety of simple auditory stimuli and the abilities to identify and to discriminate speech sounds. Although the abilities to process a battery of the simple stimuli were found to be highly correlated among themselves, these abilities did not predict performance in speech perception.

The failure of past psychoacoustics research to predict performance in speech perception tasks may be in part due to its use of very simple stimuli and concentration on the study of the processing of fine, rather than global, structure of complex auditory stimuli. It is quite possible that the strategy used to selectively listen to acoustic details differs from that used to listen to global aspects of speech and other complex nonspeech sounds. To detect details of complex auditory stimuli, listeners are instructed to focus attention on very specific stimulus properties. Results from experiments using such procedures and designed to evaluate the limits on sensory processing, obviously have made significant contributions to our understanding of the physiological mechanism of the cochlear in processing frequencies and intensities. However, to understand speech, the auditory system may not have to resolve all details in a speech signal; in fact, such detailed, focused processing might hinder more integrative, global processing. Instead, the most relevant stimulus properties and the redundancy of various cues that exist in speech sounds may require listeners to capture global properties of the sounds. Because, in a normal listening environment, the fine structure of speech sounds is seldom modified by various environmental factors or by mixing with other sounds (including different speech sounds), such "global structure" listening strategy seems to be more ecologically valid than the "fine structure" listening strategy.

In contrast to earlier work, recent psychoacoustics research has reported that listeners can use broad frequency ranges of information even when asked to detect a change in a local frequency component. Examples of such research include studies of profile analysis (e.g., Green, 1988), comodulation masking release (e.g., Hall, et al., 1984), comodulation detection difference (McFadden, 1987; Wright, 1990), modulation detection interference (Yost & Sheft; 1989; Yost, et al., 1989), and correlational listening (Cohen & Schubert, 1987). These different types of research infers global processing based upon changes in the detection or the discrimination of frequency components. The current study differs from these psychoacoustics research efforts in that it directly investigates listeners' abilities to discriminate a global spectral property that has been identified as a major factor in determining speech quality.

The global property investigated in the current study is the spectral slope of complex stimuli, with the stimuli synthesized with a simple harmonic structure on the basis of the source-filter model of speech production (Fant, 1960). Like speech, a stimulus is produced by convoluting a source spectrum with a filter transfer function. In the current study, the source spectrum is composed of the first 20 harmonic frequencies of a given fundamental frequency, with the intensities specified by a decreasing spectral envelope. Besides varying in the slope of the spectral envelope (spectral slope), the stimuli also differ in terms of a number of other properties, including the fundamental frequency (typically another property of the sound source) and the number of spectral peaks (typically a filter or resonator, rather than source, property). Listeners are asked to discriminate spectral slope while ignoring irrelevant variation in the fundamental frequencies, spectral peaks, and overall stimulus intensity. The current study thus examines the perceptual invariance of spectral slope in the context of variation in frequency composition (fundamental frequency) and in spectral shape (spectral peaks).

The choice of the three variables (fundamental frequency, resonator characteristic, and spectral slope) is motivated by the resemblance of these properties to important aspects of speech sounds. The manipulation of these three stimulus variables (dimensions) thus captures important variations observed in speech sounds. Speech produced by male and female speakers differs largely in fundamental frequency. The pattern of spectral peaks defines formant structure, and perceptually differentiates linguistic categories of speech sounds (e.g., vowels). Spectral slope, on the other hand, determines various paralinguistic characteristics of a speaker (e.g., phonation, voicing style, etc., Monsen & Engebretson, 1977; also summarized in Stevens, 1989). For example, breathy voice tends to have steeper spectral slope than modal voice, whereas creaky voice has shallower spectral slope than modal voice. In addition, spectral slope differs between speech sounds of different genders, with female voices typically exhibiting steeper spectral slope than male voices (Klatt & Klatt, 1990; Monsen & Engebretson, 1977; Price, 1989). The three stimulus dimensions therefore represent two categories of information typically carried in a complex sound (e.g., speech), with both fundamental frequency and spectral slope defining source properties, and the structure of spectral peaks defining a resonator characteristic.

The use of the source-filter model in synthesizing the stimuli and examining the relationship between the source and filter properties also is motivated by an ecological validity. Specifically, a complex sound can be viewed as a sequential operation of the convolution of a sound source and a filter transfer function; and listeners are believed to possess a highly developed ability to parse the complex sound into the source and filter transfer functions on the basis of implicit knowledge acquired in life (Gaver, 1993; Li, Logan, & Pastore, 1991; McAdams, 1993; Woods & Colburn, 1992). The most familiar examples of the application of source-filter model are in studies of vowel recognition and speaker identification. A vowel sound is produced by passing the glottal exciting source through the vocal tract which imposes formant structure on the spectrum. Listeners are not only able to use formant structure to identify the vowel category, but also able to uncover the speaker characteristics carried by the glottal source. The current study will evaluate the influence of variation in the filter property on listeners' ability to discriminate the source property, and thus investigate the perceptual relationship between the two. Gagne and Zurek (1988) examined effects of speech source upon formant discrimination, which thus represents a type of mirror image to aspects of the current study. Because a complex sound can be the joint effect of any sound source and any filter transfer function, all such studies in the extreme are bound to fail. However, by carefully choosing the source and filter transfer function, such studies should help us to understand the perceptual interaction between the properties determined by the source and filter transfer function. Thus, the current study has used typical speech parameters in defining the stimulus properties.

Although little psychophysical research has directly examined perceptual invariance or constancy for global stimulus properties of complex sounds, many speech studies have revealed a constancy in listeners' abilities to map phonetic category from speech sounds that vary significantly in waveform. Because nonlinguistic variations in speech sounds are hypothesized to be "removed" or "partitioned out" prior to phonetic identification, this type of finding has been called "normalization." The "normalization" processes typically are described as factoring out fundamental frequency (intrinsic "normalization"), the characteristics of the speaker (extrinsic "normalization"), or both. Global invariant properties such as the formant structure of vowels or the spectral shape of stop consonants (see below) are then easily extracted for identification of the phonetic category. However, listeners' abilities to "normalize" speech and other complex sounds have been explored only superficially and mechanisms for the "normalization" processes are not understood. The concept of "normalization" used in this study thus only refers to recovery of relational (sometimes also global), rather than absolute, properties in the recognition of an auditory event. Evidence for the use of global properties in speech perception was reported by Stevens and Blumstein (1981), who identified three types of global spectral shape from initial consonantal release as important cues for classification of stop consonants. Despite variability in the microstructure of consonantal spectra, Stevens and Blumstein (1978; Blumstein & Stevens, 1979) found that diffuse rising and diffuse falling spectral shape serve to cue, respectively, bilabial and alveolar consonants, and spectral shape compacted in the mid-frequency range cues velar consonants.

Speech sounds are not the only auditory event that involves the recovery of certain global properties. Music melodies, for example, are recognized by relative frequency relations rather than absolute frequencies. Changes in absolute frequencies have very little effect on listeners' abilities to identify a music melody as long as the relative frequencies among the components are preserved. Kidd and Watson (1989) used a sequence of five tones as stimuli to examine listeners' abilities to detect a frequency change in a target component when the tonal sequence was transposed in the absolute frequency, but the frequency relation among the tonal components remained constant. It was found that surprisingly small amounts of frequency transposition (1-2 semitones) led to a very large increase in thresholds. However, minimizing pattern uncertainty (by presenting the same pattern on every trial) resulted in dramatic improvements in performance, and, in some cases, the thresholds were almost comparable to those for absolute frequency detection. This study suggests that listeners are able to extract a relational frequency property, at least, from familiar patterns. In a more direct study of "normalization" of music sounds, Cho, Pastore, and Hall (1991) demonstrated that musically experienced listeners can factor out differences in instrument timbre, while perceiving chords.

The current study consisting of four experiments will evaluate the discrimination of the global property of spectral slope. The discrimination of spectral slope requires that listeners rely on the intensity relationship across the spectrum to discriminate spectral slope. Experiment 1 examines a potential confounding variable that may be used by listeners when asked to discriminate spectral slope. The remainder of the study has two major parts. The first part (Experiment 2) examines the effects on spectral slope discrimination when the frequency composition of the stimulus is changed due to variation in fundamental frequency. This design reflects a type of analog to processing speech sounds in which the speaking characteristics remain constant despite variation in the fundamental frequency. The second part of this study (Experiment 3) modifies the microstructure of the spectral envelope by imposing spectral peaks. In understanding speech, listeners need to rely on the gross spectral filter transfer function across frequencies regardless of amplitude variation in local frequencies. In identifying a voice, listeners need to rely on gross tendency in source properties (F_0 and spectral slope) regardless of the filter properties imposed by the vocal tract. Thus, the current design represents a type of investigation of the invariance of speaking characteristics in speech sounds produced by speakers despite differences in vocal tract configurations across phonemic categories. Experiment 4 explores an issue not resolved from Experiments 2 and 3.

General Method

Stimuli

The 200 ms stimuli were line spectra synthesized by digitally summing 20 harmonic sinusoidal frequencies in a sine phase. The digital stimuli were shaped by 5 ms linear onset and offset ramps. Following an A/D conversion (12-bit at a 10 Khz sample rate), the stimuli were low-pass filtered at 4 kHz (via a series of ITHACO Model 4302 filters, yielding 48 dB/octave) prior to being presented binaurally over TDH-49 P headphones to subjects in an acoustic chamber. For each stimulus, the intensity of a frequency component was specified mathematically for a given spectral envelope. For Experiments 1 and 2, (with the latter varying the F_0 as the irrelevant dimension), each stimulus had a linear flat spectral envelope with a negative spectral slope, as shown in Figure 1 a and Eq. (1),

$$I_i(i) = \text{Slope} * (i) + b, \qquad (1)$$

where $I_i(i)$ is the intensity of the ith frequency component ($i = 1$ to 20) and b is the F_0 intensity. For a flat spectral slope stimuli, b was set to the maximum level (72 dB), which produced no clipping of the waveform. Six levels of spectral slope were chosen in a range from -0.50 to -1.75 dB per frequency component in a step of -0.25 dB. An example of these flat spectrum stimuli is illustrated in Figure 1 a.

Insert Figure 1 about here

Experiment 3 used a resonator characteristic as the irrelevant dimension. The spectral envelope for each of these stimuli was determined by convoluting a flat spectral slope stimulus with a sinusoidal (spectral) resonator transfer function. We used a sine

function to simulate a formant-like characteristic. The resonator spectral envelope is specified as

$$I_r(i) = m * \sin(2ki/20), \qquad (2)$$

where $I_r(i)$ is the weighting function applied to the amplitude of the ith frequency component. This function specifies a resonator depth ($m = 4$ dB) and the number of resonator frequencies (k = the number of poles or spectral peaks). For these ripple spectrum stimuli, the F_0 intensity (b in Eq. 1) was reduced to 68 dB to avoid clipping. An example of these stimuli is shown in Figure 1b. This general type of resonator characteristic also was used by Berstein and Green (1987).

Procedure

Based upon a pilot study, an XAB discrimination task seemed to be more effective than an AX task; in fact, all subjects had a great deal of difficulty in performing the AX version of the task. On each trial, a standard stimulus (X) was always presented in the first interval, followed by the A and B test stimuli. One of the test stimuli was identical in spectral slope to the standard stimulus with the other stimulus differing in spectral slope. Subjects indicated which of the two test stimuli matched the standard stimulus in spectral slope by pressing a button on a response box. The correct response to both panels in Figure 2 should be stimulus B. (The explanation for this figure is given below.) Trial-by-trial feedback was provided following the subjects' responses. The three stimuli on each trial were separated by 500 ms. There was an 800 ms inter-trial interval.

The current study employed two types of experimental conditions to examine the effects of variations in an irrelevant dimension on the discrimination of a relevant dimension (spectral slope). On a single trial in a roving-irrelevant-dimension condition, two stimulus dimensions varied simultaneously, with listeners required to respond to the difference on the relevant dimension, but to ignore the variation on an irrelevant dimension. For this roving condition, the two test stimuli (A and B) on a single trial had the same value on an irrelevant dimension that differed from that of the standard stimulus (X). On a trial in a fixed-irrelevant-dimension condition, a single value of an irrelevant dimension was used for all three stimuli. Figure 2 shows two roving conditions which differ in the type of irrelevant dimensions: fundamental frequency (Fig. 2a) and number of spectral peaks (Fig. 2b).

Insert Figure 2 about here.

For each subject, a hit/false-alarm matrix was constructed for each experimental condition as a function of differences in spectral slope. Discrimination indices (d') were then calculated from this matrix following the ABX response model suggested by Macmillan and Creelman (1991; also Pierce & Gilbert, 1958).

Subjects

Six SUNY-Binghamton students were paid for their participation in this study. The subjects originally were naive to auditory psychoacoustics tasks and all reported normal hearing.

Experiment 1

Spectral slope was defined in terms of a systematic decrease in intensity of a frequency component with an increase in frequency. Because the current study fixed the F_0 intensity, stimuli with a steeper spectral slope had lower intensity in high frequencies relative to those with a shallower spectral slope. Thus, spectral slope co-varied with overall stimulus intensity (and intensity in the upper portion of the spectrum. By definition, the manipulation of spectral slope must co-vary with the absolute intensity of some portions of the stimulus spectrum). As a result, instead of judging the global property of spectral slope, the subjects might be able to perform the task by comparing either overall stimulus intensity or intensity in a specific high frequency region. Experiment 1 explored the possible use of either of these alternative listening strategies in the spectral slope discrimination task. Two experimental conditions were used. In the fixed-level condition, the F_0 intensity was equal for the three stimuli on a trial, and therefore, overall stimulus level was determined solely by spectral slope (see Eq. 1). In the roving-level condition, F_0 intensity was determined randomly and independently over a range of 20 dB for each of the three stimuli on a trial, thus precluding an effective use of the absolute level in performing the spectral slope discrimination task. This roving of intensity procedure has been used in profile analysis studies to eliminate responding to the absolute level (Green, 1988; Mason, Kidd, Hanna, & Green, 1984).

There are three possible patterns of outcomes which could occur with the two conditions. First, if subjects use the overall stimulus level as a cue, but cannot respond directly to spectral slope, discrimination performance should be high in the fixed-level condition, and at chance in the roving-level condition. This finding would suggest that the current project should be discontinued. Second, if subjects use information only about spectral slope, and thus, do not use the correlated intensity cue, equivalent discrimination performance should be observed in the two conditions, and subsequent experiments then need not control for differences in overall stimulus level caused by varying spectral slope. Finally, if listeners use both types of information, discrimination performance should be higher in the fixed-level condition than in the roving-level condition, with performance for both conditions above chance. The last possible outcome will require roving overall stimulus level to eliminate this correlated intensity cue.

Stimuli and Procedure

Six stimuli were synthesized by using the six levels of spectral slope with a 170 Hz fundamental frequency. For the roving-level

condition, the overall intensity of each stimulus within a trial was determined randomly in a range of 20 dB in 1 dB steps, and thus the I0 intensity was ranged from 52 to 72 dB. Roving overall stimulus intensity was implemented using a Charybdis Model D programmable attenuator. Subjects completed 12 blocks of trials. The 60 trials in a block were created by crossing each of the six levels of spectral slope with every other level and with equal assignment of the paired spectral slope values as the standard stimulus. The data collection began after 720 practice trials.

Results and Discussion

Figure 3 shows the individual discrimination performance for the two conditions as the function of slope differences. The solid curves with open circles show the results of the fixed-level condition, and the dotted curves with filled circles are those of the roving-level conditions. These curves are essentially psychometric functions, plotting the discrimination ability as a function of the magnitude of spectral slope difference. However, they differ from more standard psychometric functions in that each point along the abscissa represents the pooling of all stimulus pairs that differ by the given magnitude of spectral slope. The fixed-level condition for all subjects, and the roving-level condition for some subjects, reaches ceiling ($d' > 6.0$) at the largest slope differences. In fact, even d' values of 5.0 or greater cannot be very accurate since all differences in z scores reflect extremely small changes in the upper 5 % of the tail of the normal distribution. Moreover, the d' values for the larger slope differences were based upon fewer stimulus pairs that differed by the given magnitude than those for the smaller slope difference. (Experiment 4 will examine the importance of unequal sampling.)

Insert Figure 3 about here.

With the exception of Subject 2, all subjects exhibited linear psychometric functions for both fixed- and roving-level conditions, but with a shallower slope for the roving-level condition. The shallower slope for the roving-level condition than for the fixed-level condition indicates that the subjects were all using the absolute stimulus intensity information in the latter condition, but could still perform the spectral slope discrimination task when the absolute intensity provided no valid information. Subject 2 was not able to perform the roving discrimination task to a reasonable level (e.g., $d' > 1$) at even the largest difference in spectral slope, indicating that this subject was probably using only absolute intensity to perform the discrimination.[1] The average psychometric functions (Figure 3g) thus are based upon the five subjects who could perform the discrimination task under the roving condition (i.e., excluding Subject 2).

Both mean psychometric functions are highly linear before reaching ceiling with essentially zero intercepts ($r^2 = 0.99$ for both conditions). The two functions differ solely in the slope of the linear regression equations (7.6 for the fixed- and 4.9 for the roving-level condition). The ratio of these values is 1.55, which can be interpreted as the absolute intensity cue contributing approximately 35 % of decision information to the fixed discrimination. An alternative interpretation of this ratio (adopting the formula from Macmillan, Braida, & Goldberg, 1987) is that this roving intensity information increases the variability of perceptual decision, with the ratio of variances being 1.40;[2] this interpretation is more appropriate for the later experiments where the roving variable clearly functions only by adding noise, rather than by also eliminating a correlated cue for discrimination.

A 2 x 5 within-subject analysis of variance was performed on the d' values. The significant main effect of the spectral slope difference confirmed that discrimination improved as the spectral slope difference was increased ($F(4,20) = 62.90$, $p < .05$), as exhibited by the linear psychometric functions. The significant main effect of the conditions indicated better discrimination performance for the fixed-level condition than for the roving-level condition ($F(1,5) = 33.96$, $p < .05$). The interaction of the (fixed- and roving-level) conditions and the spectral slope difference was not significant ($F(4,20) = 2.07$, $p > .05$).

For the roving-level condition, discrimination performance was clearly above chance. To appreciate this finding, the reader is reminded that, on each trial in the roving-level condition, overall stimulus level was varied randomly across the three stimuli; therefore the stimulus with the steepest spectral slope might have higher overall level than those with shallower spectral slopes. It thus was not possible for the subjects to make accurate judgment on the basis of overall stimulus level (or absolute intensity in any portion of the spectrum). This result indicates that the five subjects had to extract some sort of sound quality based on spectral slope from these complex stimuli. However, compared with that in the fixed-level condition, discrimination performance was poorer in the roving-level condition. This latter result indicates that all subjects exploited overall stimulus level as an additional cue for the difference in spectral slope, and thus requires that subsequent experiments eliminate the intensity cue in evaluating spectral slope discrimination. Therefore, in the subsequent experiments, overall stimulus lev[illegible] always be varied randomly over a range of 20 dB within a trial. (The fixed versus roving conditions in Experiments 2-4 will differ in terms of an irrelevant stimulus dimension, with both conditions in each experiment roving overall stimulus level.)

Experiment 2

Experiment 2 investigates the effects of variation in the fundamental frequency on spectral slope discrimination. Both fundamental frequency and spectral slope are considered as two sound source properties; it is quite possible that listeners will treat all sound source properties as a unit in a sound-producing system. The two conditions in this experiment were used to examine the listener's ability to extract spectral slope. In the first condition, fundamental frequency was fixed within a trial, but randomly varied across trials (the fixed condition). In the second condition, fundamental frequency was randomly varied both within a trial and across trials (the roving condition). The performance difference between the two conditions will indicate the magnitude of the adverse effects

of variation in the fundamental frequency upon spectral slope discrimination, and therefore, should shed light upon the perceptual relationship between the two source properties. The research method is similar to what was used by Durlach, Tan, Macmillan, Rabinowitz, and Braida (1989), and represents an accuracy-based version of the Garner (1974) strategy for evaluating integral versus separable dimensions.

Stimuli and Procedure

The six levels of spectral slope from Experiment 1 were combined with three levels of fundamental frequencies (150, 170, and 190 Hz) to create 18 stimuli. These specific fundamental frequencies were chosen to guarantee no common frequency components for the stimuli differing in the fundamental frequency. In the fixed condition, the three stimuli on each trial shared a common fundamental frequency, equally selected from the three fundamental frequencies. In the roving condition, two different fundamental frequencies were selected for the standard stimulus and the two test stimuli on each trial. There were six possible combinations of fundamental frequency pairs by crossing the three fundamental frequencies with every other. For each condition, the subjects completed 12 blocks of 60 trials, always preceded by 720 practice trials. (Because of roving intensity under all conditions, the fixed condition with the 170-Hz F_0 is identical to the roving-level condition in Experiment 1.)

Results and Discussion

Figure 4 shows discrimination performance for each subject under the two conditions in which the fundamental frequency was fixed (solid curves with open circles) or randomly varied within a trial (dotted curves with closed circles). With the exception of Subject 2 under the roving condition, the subjects under both conditions exhibited a generally linear growth in discrimination as a function of an increase in the spectral slope difference. Figure 4 g describes the mean d's across the five subjects (excluding Subject 2). For d' values under 6.0, the two mean functions are linear (r^2 = 0.99 and 0.98 for fixed and roving conditions, respectively), with intercepts of -0.13 and -0.25. These results were confirmed by the presence of a significant main effect of the spectral slope difference in a 2 x 5 within-subject analysis of variance ($F(4,20) = 62.51$, $p < .05$). The results clearly indicate that the subjects were able to evaluate spectral slope despite variation in the fundamental frequency. Consider the roving condition in which the stimuli differed in the frequency components due to the change in the fundamental frequency. Across both conditions in this experiment, the only relational property that remains constant regardless of variation in the fundamental frequency (and in overall stimulus intensity) is spectral slope. Therefore the subjects had to rely on the relational (or global) properties defining spectral slope to make correct responses.

Insert Figure 4 about here.

Poorer discrimination performance was found in the roving fundamental frequency condition compared with the fixed fundamental frequency condition ($F(1,5) = 26.145$, $p < .05$). The adverse effect due to the variation in the fundamental frequency was evident across all the six subjects, indicating that the decision on spectral slope was influenced to some degree by the difference in the fundamental frequency.[3] There was a nearly significant interaction between the levels of the spectral difference and the two conditions ($F(4,20) = 2.57$, $p < .06$). Ignoring the highest levels of performance, the detrimental effect of roving fundamental frequency was in terms of a change in the slope of the psychometric function (4.4 versus 3.6 for the fixed and roving conditions), which is consistent with added noise due to the roving condition. Substituting the slopes of the regression equations for d' in an adapted version of the formula in Macmillan, Braida and Goldberg (1987), it is estimated that the ratio of variance due to roving fundamental frequency relative to variance associated with spectral slope discrimination (with roving overall intensity) is 0.49. Thus, for the specific range of values in this experiment, roving fundamental frequency adds approximately 50 % more variability to the decision process.

The results of this experiment indicate a moderate level (50 % variance increase) of Garner interference which is said to occur when slower response time or lower accuracy is found for the stimuli with incongruent values on paired dimensions relative to stimuli with congruent values (Garner, 1974). Garner interference indicates a failure of selective attention to a relevant dimension due to variation in an irrelevant dimension; thus the paired dimensions are considered as an integral perceptual unit. Because, in typical studies of Garner interference, researchers use highly discriminable binary values on paired stimulus dimensions, results from typical studies cannot be used to address the extent to which the perception of one dimension is influenced by the other. Rather researchers typically tend to draw more absolute conclusions about whether one stimulus dimension is perceptually accessible. The use of multiple values on both relevant and irrelevant dimensions in the current study has allowed a more detailed evaluation of degree to which the relation between spectral slope and the fundamental frequency is separable from the fundamental frequency.

Experiment 3

In contrast to the use of two sound source properties (spectral slope and fundamental frequency) in Experiment 2, Experiment 3 investigates the perceptual constancy of spectral slope in the context of variation in the number of spectral peaks (which defines a resonator characteristic). Because spectral slope and the resonator characteristic reflect different sound-producing components, the detrimental effect of varying the irrelevant dimension may not occur in Experiment 3. Again, two conditions were used to determine the effect of roving the resonator characteristic on spectral slope discrimination. In the fixed condition, the number of spectral peaks was fixed within a trial, but varied across trials. In the roving condition, the number of spectral peaks was varied both within a trial and across trials.

Stimuli and Procedure

The six levels of spectral slope (from Experiment 1) and three levels of the number of spectral peaks (2, 5, and 8 peaks) were combined to create 18 stimuli with a 170 Hz fundamental frequency. Spectral peaks were equally spaced across the frequency range. The number of spectral peaks was determined in a pilot study in which the subjects discriminated stimuli generated by convoluting a flat spectral envelope (with a zero spectral slope) with the filter transfer functions differing in the number of spectral peaks. The subjects showed above 90% accuracy even when discriminating stimuli differing by only one spectral peak. Therefore, stimuli that differed by three spectral peaks on a trial in the roving condition are highly discriminable from each other, and should cause a significant perceptual variation for the stimuli. The question is whether this significant perceptual variation has any effect on spectral slope discrimination. A block of 60 trials were created using the pairing strategy adopted in Experiment 2 (including the roving of overall spectral intensity). The procedure for roving the number of spectral peaks was identical to that for roving fundamental frequency in Experiment 2. For each condition, the subjects completed 720 practice trials, then 12 blocks of 60 trials.

Results and Discussion

Figure 5 shows the individual performance and the mean discrimination across the five subjects (excluding Subject 2) for each condition as a function of the magnitude of the slope difference. With the usual exception of Subject 2 under the roving condition, subjects showed spectral slope discrimination improved with an increasing difference in slope, with nearly identical performance under both conditions. Figure 5 g is the mean psychometric function (again excluding Subject 2). Consistent with the findings from Experiment 2, the observation that performance improved with an increase in the slope difference was confirmed by the presence of a main effect of spectral slope difference ($F(4,20) = 58.458$, $p < .000$). The linear growth in d' (for $d' < 6.0$) with the increasing slope difference ($r^2 = .99$ for both conditions) indicated that the subjects again were able to use spectral slope as a cue to differentiate different degrees of spectral slope.

Insert Figure 5 about here

Although overall discrimination performance in the fixed condition was slightly better than that in the roving condition ($F(1,5) = 13.897$, $p < .013$ based upon data from all six subjects), the planned comparison test showed that the significant statistical difference occurred only for the stimulus pair with the largest slope difference, and was mostly due to the failure of Subject 2 in performing this (or any) roving task. The mean psychometric functions (excluding Subject 2) for $d' < 6.0$ are essentially identical in slope (4.9 vs. 5.0), and intercept (-0.13 vs. -0.41) for the fixed and roving conditions, respectively. Following Macmillan, Braida and Goldberg (1987; see Footnote 1), virtually no additional variance was added into the subject's judgment by roving the number of spectral peaks relative to the fixed condition. The lack of this roving effect is in sharp contrast with the significant difference across most levels of the spectral difference in Experiment 2, and thus suggests that spectral slope discrimination is independent of the resonator characteristic (at least when defined in terms of the number of spectral peaks).

In this experiment, listeners had to extract a global spectral property of slope on the basis of overall spectral tendency, while "tolerating" variation in the fine structure of spectral envelope. Although this result was obtained from nonspeech complex stimuli, it is consistent with those from speech sounds. For example, Stevens and Blumstein (1981) reported that, despite variations in spectra of initial consonant release, listeners used overall tendency of spectral shape as a cue to classify stop consonants contrasted in place of articulation. This result also parallels that of Gagne and Zurek (1988) who, measuring thresholds for formant discrimination when varying different types of speech source (i.e., harmonic series and broadband noise), showed that listeners were able to discriminate formant (thus resonant) frequency equally well for the harmonic and noise glottal source. Although the two studies focused on the opposite stimulus property (with the current study examining source discrimination in the context of variation in the resonator characteristic), the two studies seemed to agree that sound source and resonator characteristics can be processed independently. In a different type of study of sound source perception, Freed (1990) asked listeners to judge the hardness of the mallet when it struck different types of pans, with the mallet and the pan, respectively, serving as sound source and resonator. Freed found that listeners were able to judge the hardness of the mallet independent of types of pans. This finding is consistent with the results reported here.

Experiment 4

Experiment 4 was designed to address two unresolved issues from the preceding experiments. First, a lack of performance difference between the roving and the fixed conditions was found in Experiment 3 (varying spectral peaks) in contrast to a significant difference in Experiment 2 (varying fundamental frequency). Because Experiment 2 preceded Experiment 3, the difference in roving task performance might be explained in terms of differences in experience with the task. Experiment 4 examined this potential confound by partially replicating Experiment 2 using a sample of the stimulus pairs. Additionally, because the current study computed d's as a function of levels of slope difference for all stimulus pairs with a given magnitude of the difference, the d' values for the larger slope difference were based upon a fewer stimulus pairs than the d's for the smaller slope difference (i.e., for the range of slopes between 0.50 and 4.75 dB in steps of 0.25, there are 5 comparisons differing by 0.25 dB, 4 differing differ by 0.50 dB, and only 1 differing by 4.25 dB). Experiment 4 examined whether this unequal sampling of slope difference may have influenced the stability of the d' values.

Stimuli and Procedure

The 18 stimuli from Experiment 2 were used to generate five pairs of stimuli with the slope difference listed in Table 1. The selection of the stimulus pairs was guided by (1) having each stimulus pair represent one level of the slope difference; and (2) having each level of the slope difference occur with an approximately equal probability. These five pairs of stimuli were randomly presented ten times in a block of 50 trials. For each condition, the six subjects completed six blocks of trials in a one-hour session. The sequence of running the two conditions was counterbalanced across the subjects.

Insert Table 1 about here

Results and Discussion

The results from this replication of Experiment 2 were quite similar to those found originally in Experiment 2. A 2 x 5 within-subject analysis of variance was conducted on the discrimination performance. The significant main effect of the slope difference again indicated that discrimination performance differed among the levels of slope differences ($F(4,20) = 39.287$, $p < .05$). The main effect of conditions showed that discrimination performance was lower significantly in the roving condition than in the fixed condition ($F(1,5) = 60.406$, $p < .05$). Figure 6 shows a significant difference in the d's for the largest slope difference (stimulus pair 5 from Table 1) between the fixed and the roving conditions. The similarity between the psychometric functions of Experiments 2 and 4 indicates that the results of these preceding experiments had not been significantly biased due to unequal samples of data used to compute the d's.

Insert Figure 6 about here

The new results for the fixed discrimination condition yielded a linear psychometric function ($r^2 = 0.98$), with a slope of 4.9, which is identical to the psychometric functions for the equivalent conditions in Experiments 1 and 3. The slightly depressed level of performance found in the original fixed condition (Experiment 2) thus would seem to be attributable to a relative lack of experience with the roving task. The roving condition also exhibited a linear psychometric function ($r^2 = 0.99$), but with a somewhat shallower slope than the psychometric function obtained in Experiment 2 (2.9 versus 3.6). Because the total range of spectral slopes sampled in Experiments 2 and 4 was identical (-0.50 to -1.75 dB per harmonic), the difference in slope of the psychometric functions cannot be attributable to a simple range-related difference in context variance across the two tasks. However, it is quite possible that subjects were bothered more by the increase in frequency of a larger change in the irrelevant dimension (e.g., a change in spectral slope of -1.25 occurred with a provability of 0.07 in Experiment 2 and a probability of 0.20 in Experiment 4). The persistence of this large detrimental effect of variation in the fundamental frequency suggests that the absence of any decrement in Experiment 3 cannot be attributed to the practice. The results of Experiment 4 thus strengthen the speculation that spectral slope is more easily separated from the resonator characteristic than from the fundamental frequency.

The finding that perception of sound source is relatively independent of the resonator property becomes more compelling when examining the relevant results across the four experiments. Figure 7 re-plots the mean discrimination performance for the fixed conditions (with a fixed irrelevant dimension, but roving amplitude) and the mean discrimination performance for the roving conditions (with a roving irrelevant dimension) from Experiments 2 (roving F_0) and 3 (roving filter function). Figure 7 shows that the magnitude of spectral slope discrimination was unaffected by varying the number of spectral peaks (Experiment 3) as an irrelevant dimension, but was significantly influenced by varying the fundamental frequency (Experiment 2 and also Experiment 4).

Insert Figure 7 about here

One account for this asymmetric result can be formulated in terms of the sound production model. The current study did not arbitrarily choose two acoustic variables as the irrelevant dimensions. Rather Experiment 2 used a definite sound source property--the fundamental frequency--as the irrelevant dimension, whereas Experiment 3 used a resonator characteristic--the number of spectral peaks--as the irrelevant dimension. Speech research has shown that many speech glottal source properties are determined by both spectral slope and fundamental frequency together. For example, various types of vocal registers differ in the fundamental frequency and spectral slope (Childers & Lee, 1991; Hollien, 1974). Moreover, speech produced by male and female speakers also differ in both fundamental frequency and spectral slope. The various speaking characteristics are thus determined by a set of congruent values of glottal acoustic dimensions. It is unlikely that incongruent values of glottal acoustic dimensions will invoke an unambiguous perception of the speaker identity. Therefore, listeners may treat all sound source properties as a perceptual unit that, as a whole, describes the characteristics of the sound production system. However, analyzing the function of sound source and filter presents a different picture. For speech sounds, different linguistic messages are generated by maneuvering the vocal apparatus (e.g., tongue, jaw, lips, teeth, etc), which, in turn, alters the resonance or filter properties. Regardless of what linguistic messages a speaker produces (by changing the filter properties), the speaking characteristics of the source always remain constant. This relative independence of the glottal source and vocal tract configuration may dictate the absence of the effect of variation in the number of

spectral peaks on spectral slope discrimination.

General Discussion

The current study was an attempt to utilize a psychophysical procedure to investigate the perception of complex global stimulus attributes which would be neglected in the past psychoacoustics research. Evidence has suggested that the research that focuses on detailed acoustic properties may not be able to successfully predict performance in perception of speech and other complex naturally-occurring sounds.

Because the past psychoacoustics research focuses on the abilities of auditory peripheral system to detect or discriminate changes in elementary acoustic properties such as, pitch, loudness, phase and duration of a signal, it is called the Fourier analysis approach (Gaver, 1993). Understanding listeners' abilities to process elementary acoustic properties has taught researchers a great deal about elemental aspects of perception and about the physiological underpinnings of auditory information processing. However, our auditory system also needs to recover more global stimulus properties, such as source characteristics of a sound-producing system, through listening to the sound produced by the system (e.g., Freed, 1990; Gaver, 1993; Li, et al., 1992; Repp, 1987; Warren & Verbrugge, 1984). It is doubtful that studying the processing of elementary acoustic properties is sufficient to promote a full understanding of perception of naturally-occurring sounds. Although we do not agree with the extreme position that studying the processing of elementary properties in the hope to understand perception of naturally-occurring sounds is as distant as studying the processing of features of letters in the hope to understand reading comprehension (Gaver, 1993), there clearly is merit to moving to a higher level of analysis when trying to understand more global aspects of perception.

The source-filter analysis approach, used in the current study, provides one alternative to the Fourier analysis approach. Rather than using elementary acoustic properties as stimulus dimensions, the source filter approach assumes that a complex sound, particularly, a naturally occurring sound, is produced by an interaction of sound producing components such as power, oscillator, resonator, and coupler. A sound is produced by vibrating the oscillator, which is excited by the power. The sound is then shaped by the resonator where the spectrum of the source is tuned to the resonator characteristics. Thus, the acoustic (spectral and temporal) effects of these sound-producing components should be stimulus dimensions in a laboratory study of perception of naturally occurring sounds. Although the source filter model describes sound production, the model also provides a guide for investigating perception of characteristics of sound-producing components. To implement this research scheme, two types of perceptual questions must be asked: (1) whether the acoustic effect of each component is perceptually accessible; and (2) what is the nature of perceptual interaction among the components. To investigate the first question, the physics of each component must be understood, and the acoustic effect of each component should be then examined by varying different physical parameters defining the component. Because not every acoustic effect is perceptually meaningful, a perceptual investigation needs to follow the acoustic analysis. Psychophysics analysis of the acoustic effect can help determine the mapping from the physical parameter of the sound-producing component to the acoustic effect, and then to the perceptual effect.

The first question focuses on individual components of a sound-producing system, including. The current study attempted to answer the second question of how the sound-producing components may be interacted to influence our perception of source characteristics. We varied two types of source properties and a filter property with the intent to understand the perceptual interaction between the source and filter properties. Relative to the impact exerted by the fundamental frequency (another source property), the filter property had less influence upon the perceptual resolution (discrimination) of spectral slope (a source property). The results supported the claim that the source and filter property were relatively independent (at least for the types of source and filter properties used in this study). However, it certainly remains unclear whether this finding can be generalized to the combination of any source and filter properties. Although it is likely that the validity of this finding depends on the type of source and filter properties (i.e., transfer function) used in this study, the general research approach is certainly useful for studying the perception of complex sounds, especially, naturally-occurring sounds. For example in studying the gender judgment by listening to the sounds of human footsteps (Li, et al., 1992), this approach can be used to determine what proportions of shoe (source) and walking surface (resonator) factors contribute to the gender judgment of a walker. This research is now underway in the Pastore's laboratory at SUNY-Binghamton.

Although this discussion emphasizes the importance and necessity of the source-filter analysis approach, both the Fourier analysis and the source-filter analysis approaches represent important, but different levels of analysis in investigation of human audition. In vision, Marr (1982) proposed three different levels of data structures: 1) the primal sketch, 2) the 2-D sketch, and 3) the 3-D model. An audition analog of Marr's theory was developed by Richards (1988). The primal sketch is the waveform or spectrum representation of the signal. The Fourier analysis approach certainly helps us to understand the resolution of various acoustic properties in such a representation. While the 2-D sketch consists of "visible" surface properties in vision, in audition, the 2-D sketch may include sound localization and properties of source and filter (e.g., harmonicity source, noise source; cylindrical resonator and drum; metal, wood, and glass material). The source-filter analysis approach can be used to investigate the separation of source and filter properties as well as the perceptual relationship among them. Finally, the 3-D model is a vivid and coherent description of the sound in a 3-dimensional space. This gross theoretical scheme may provide a direction for us to eventually understand the perception of complex naturally-occurring sounds, and to organize what we have known about the perception of complex sounds. The specificity and the validity of this research scheme can be tested only by more empirical research and theoretical advancement.

References

Bernstein, L. R., & Green, D. M. (1987). Detection of simple and complex changes of spectral shape. Journal of the Acoustical Society of America, 82, 1587-1592.

Blumstein, S. E., & Stevens, K. N. (1979). Perceptual invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America, 66, 1001-1017.

Childers, D. G., & Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception. Journal of the Acoustical Society of America, 90, 2394-2410.

Cho, J. L., Hall, M. D., & Pastore, R. E. (1991). Normalization process in thehuman auditory system. Journal of the Acoustical Society of America, 89, pt. 2, 1988.

Christopherson, L. A. (1992). Some psychometric properties of the Test of Basic Auditory Capabilities (TBAC). Journal of Speech and Hearing Research, 35, 929-935.

Cohen, M. E., & Schubert, E. D. (1987). The effect of cross-spectrum correlation on the detectability of a noise band. Journal of the Acoustical Society of America, 81, 721-723.

Durlach, N. I., Tan, H. Z., Macmillan, N. A., Rabinowitz, W. M., & Braida, L. D. (1989). Resolution in one dimension with random variations in background dimensions. Perception & Psychophysics, 46, 293-296.

Fant, G. (1960). **Acoustic Theory of Speech Production**. Mouton: The Hague.

Freed, D. J. (1990). Auditory correlates of perceived mallet hardness for a set of recorded percussive sound events. Journal of the Acoustical Society of America, 87, 311-322.

Gagne, J., & Zurek, P. M. (1988). Resonance-frequency discrimination. Journal of the Acoustical Society of America, 83, 2293-2299.

Garner, W. R. (1974). **The Processing of Information and Structure**. Potomac, MD: Erlbaum.

Gaver, W. (1993). What in the world do we hear? An ecological approach to auditory source perception. Ecological Psychology, 5, 1-29.

Green, D. M. (1988). **Profile Analysis: Auditory Intensity Discrimination**. NY: Oxford.

Hall, J. W., Haggard, M. P., & Fernandes, M. A. (1984). Detection in noise by spectro-temporal pattern analysis. Journal of the Acoustical Society of America, 76, 50-56.

Hollien, H. (1974). On vocal register. Journal of Phonetics, 2, 125-143.

Kidd, G. R., & Watson, C. S. (1989). Detection of relative-frequency changes in tonal patterns. Journal of the Acoustical Society of America, 86, S121.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America, 87, 820-857.

Li, X., Logan, R. J., & Pastore, R. E. (1991). Perception of acoustic source characteristics: Walking sounds. Journal of the Acoustical Society of America, 90, 3036-3049.

Macmillan, N. A., Braida, L. D., & Goldberg, R. F. (1987). Central and peripheral processes in the perception of speech and nonspeech sounds. In M. E. H. Schouten (ed.), **The Psychophysics of Speech Perception**. The Hague: Nijhoff.

Macmillan, N. A., & Creelman, C. D. (1990). **Detection Theory. A User's Guide**. Cambridge: Cambridge.

Marr, D. (1982). **Vision**. San Francisco: Freeman.

Mason, C. R., Kidd, G., Jr., Hanna, T. E., & Green, D. M. (1984). Profile analysis and level variation. Hearing Research, 13, 269-275.

McAdams, S. (1993). Recognition of sound sources and events. In S. McAdams and E. Bigand (eds.), **Thinking in Sound: The Cognitive Psychology of Human Audition**. NY: Oxford.

McFadden, D. (1987). Comodulation detection differences using noise-band signals. Journal of the Acoustical Society of America, 81, 1519-1527.

Monsen, R. B., & Engebretson, A. M. (1977). Study of variations in the male and female glottal wave. Journal of the Acoustical Society of America, 62, 981-993.

Pastore, R. E. (1981). Possible psychoacoustic factors in speech perception. In P. D. Eimas and J. Miller (Eds), **Perspectives on the Study of Speech**. Hillsdale, NJ: Lawrence Erlbaum.

Pierce, J. R., & Gilbert, E. N. (1958). On AX and ABX limens. Journal of the Acoustical Society of America, 30, 593-595.

Price, D. G. (1989). Male and female voice source characteristics: Inverse filtering results. Speech Communication, 8, 261-277.

Repp, B. H. (1987). The sound of two hands clapping: An exploratory study. Journal of the Acoustical Society of America, 81, 1100-1110.

Richards, W. (1988). Sound interpretation. In W. Richards, (ed.), **Natural Computation**. Cambridge, MA: MIT.

Stevens, K. N. (1989). On the quantal nature of speech. Journal of Phonetics, 17, 3-45.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 64, 1358-1365.

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas and J. Miller (Eds), **Perspectives on the Study of Speech**. Hillsdale, NJ: Lawrence Erlbaum.

Warren, H., & Verbrugge, R. R. (1984). Auditory perception of breaking and bouncing events: A case study in ecological acoustics. Journal of Experimental Psychology: Human Perception and Performance, 10, 704-712.

Watson, C. S. (1991). Auditory perceptual learning and the cochlear implant. The American Journal of Otology, Supplement.

Watson, C. S., Qiu, W. W., Chamberlain, M. M., & Li, X. (1993). Auditory and visual speech perception: Confirmation of a modality-independent source of individual differences. Journal of the Acoustical Society of America. (under review)

Wright, B. A. (1990). Comodulation detection differences with multiple signal bands. Journal of the Acoustical Society of America, 87, 292-303.

Woods, W. S., & Colburn, H. S. (1992). Test of a model of auditory object formation using intensity and interaural time difference discrimination. Journal of the Acoustical Society of America, 91, 2894-2902.

Yost, W. A., & Sheft, S. (1989). Across-critical-band processing of amplitude-modulated tones. Journal of the Acoustical Society of America, 85, 848-857.

Yost, W. A., Sheft, S., & Opie, J. (1989). Modulation interference in detection and discrimination of amplitude modulation. Journal of the Acoustical Society of America, 86, 2138-2147.

Author Note

This research was supported inpart by the NSF and AFOSR grants awarded to Richard E. Pastore, and facilitated by the assistance from the Center for Cognitive and Psycholinguistic Sciences at the State University of New York at Binghamton. The preparation of this manuscript by the first author was funded by the NIH and AFOSR grants awarded to Charles S. Watson at Indiana University. Any opinions, findings, and conclusions expressed in this publication are those of the authors and do not reflect the views of the funding agencies. The first author is currently affiliated with the Center for Speech Processing, the Department of Electrical and Computer Engineering, The Johns Hopkins University. Requests for reprints should be sent to Xiaofeng Li, The Center for Speech Processing, the Department of Electrical and Computer Engineering, The Johns Hopkins University, Baltimore, MD 21218.

Footnotes

[1] In all of the experiments, only Subject 2 failed to exhibit reasonable levels of performance under the roving conditions. The fixed condition of Experiment 2 is essentially equivalent to the roving-level condition in Experiment 1, yet the performance of Subject 2 was equivalent to that of the other subjects. Therefore, it would appear that Subject 2 failed to understand the requirements of the roving conditions.

[2] Macmillan, Braida and Goldberg (1987) used the following formulae to evaluate the relative variance contributed by sensory and context codings,

$$\text{Context Variance/ Sensory Variance} = (\underline{d}'_{fixed}/\underline{d}'_{roving})2 - 1. \qquad (3)$$

We substituted the ratio of $\underline{d}'$ slopes from the fixed and roving conditions for the $\underline{d}'$ ratio in this formulae to evaluate the additional variance introduced by roving an irrelevant dimension.

[3] In this research, we have defined spectral slope in terms of change in dB per harmonic of the fundamental frequency. This definition makes a great deal of sense in terms of a production system which differs in terms of the fundamental frequency of a glottal source. This definition of spectral slope also means that the spectral envelope is maintained across F_0 when defined as a function of octave, or any other behaviorally- or physiologically-relevant unit (usually a logarithmic function of frequency). Thus, a change in the frequency region of the signal, accomplished by shifting F_0, should result in a maintaining equivalent changes in intensity per octave or critical band as a function of increasing frequency. Although changes in both F_0 and ripple (filter) frequency will alter the absolute amplitude of energy in any given frequency region (defined in terms of either fixed or behaviorally-relevant units), the roving of amplitude under all conditions makes absolute amplitude cues irrelevant to the discrimination task. However, if one believes that spectral slope is better defined in absolute physical units (change in dB per fixed unit of frequency; for example, average change in dB per Hz), then F_0 and spectral slope are related variables, whereas ripple frequency (defined across the full spectrum of the signal) and spectral slope are independent variables. Because we are evaluating perception, and the applicability of the source-filter model to perception, the definition of terms in behaviorally-relevant terms is most logical strategy.

Although future research might try to eliminate the confound (when defined in physical terms) between F_0 and spectral slope, the task is not simple. For example, defining spectral slope in terms of a constant change in amplitude per unit Hz, by its very nature, forces a spectral slope to be correlated with frequency when interpreted in terms of any of the more behaviorally-relevant definitions of slope (e.g., dB per critical band). Therefore, running the same task with this alternative definition of spectral slope should result in lower performance in the fixed condition (due to behaviorally nonlinear nature of the resulting spectral slope) and an even greater drop in performance under the roving F_0 condition (due to the correlated change in behaviorally-relevant definition of slope). We thus believe that we have evaluated the relevance of two potentially independent perceptual variables on the perception of a single perceptual variable.

Table 1. The Values of Spectral Slope for Stimulus Pairs Presented on a Single Trial for Experiment 4. Slope 1 and Slope 2 were randomly assigned to the standard and test stimuli on a trial.

Stimulus Pairs	Slope 1	Slope 2	Slope
1	-1.00	-1.25	0.25
2	-0.75	-1.25	0.50
3	-0.75	-1.50	0.75
4	-0.50	-1.50	1.00
5	-0.50	-1.75	1.25

Figure Captions

Figure 1. Schematic representation of the stimuli. The panel consists of a 20 harmonically-related complex sound (shown by the vertical lines) with a decreasing spectral envelope. The abscissa is the harmonic number, and the ordinate is the spectral intensity. The bottom panel illustrates a stimulus produced by passing the stimulus in the top panel through a five spectral peak filter.

Figure 2. Illustration of the trial structure for Experiments 2, 3 and 4. The top panel is the trial structure for Experiment 2, whereas the bottom panel is for Experiment 3. For each stimulus shown by a schematic spectrum, the vertical lines are the 20 harmonic tones with the spectral intensity specified in the ordinate. For the detailed description of the trial structure, see the text.

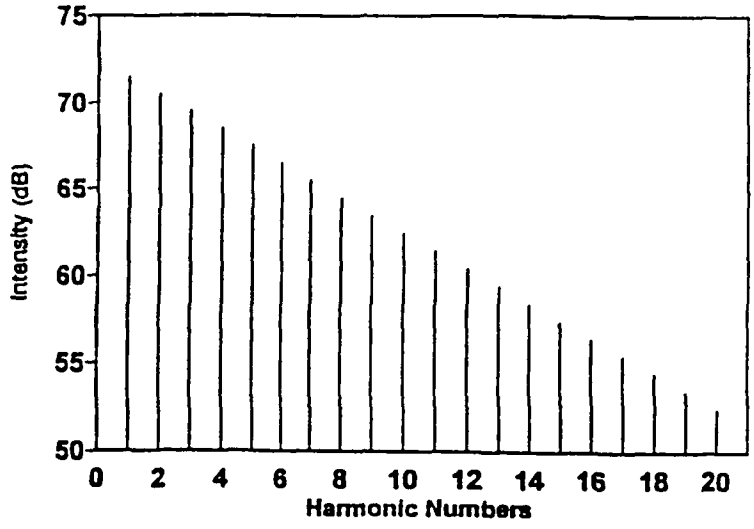
Figure 3. Results for Experiment 1. The first six graphs are the results for the individual subjects. The last graph is the mean across the five subjects (excluding Subject 2). For each graph, the abscissa is the difference in spectral slope within a trial, and the ordinate is the d' value. The solid curves with open circles are the results for the fixed-level condition, and the dotted curves with closed circles are those for the roving-level condition.

Figure 4. Results for Experiment 2. The solid curves with open circles are the results for the fixed irrelevantdimension (fundamental frequency) condition, and the dotted curves with closed circles are those for the roving irrelevant dimension condition.

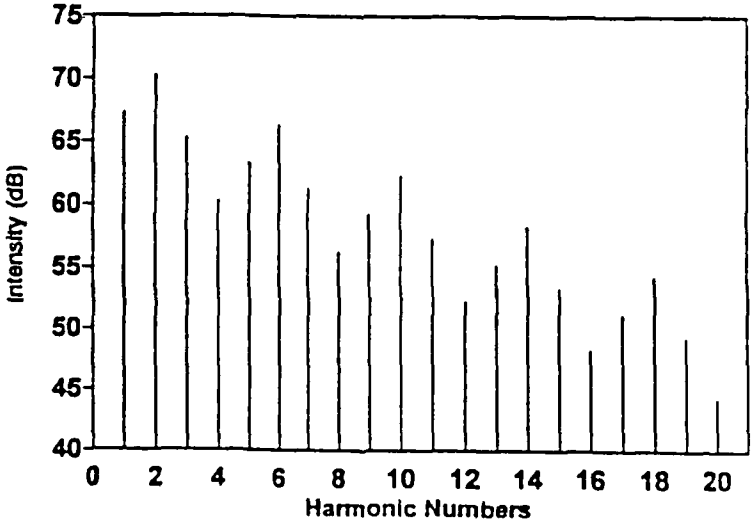
Figure 5. Results for Experiment 3. The solid curves with open circles are the results for the fixed irrelevantdimension (the number of spectral peaks) condition, and the dotted curves with closed circles are those for the roving irrelevant dimension condition.

Figure 6. Results for the replication of Experiment 2. The solid curves with open circles are the results for the fixed irrelevantdimension (fundamental frequency) condition, and the dotted curves with closed circles are those for the roving irrelevant dimension condition.

Figure 7. Comparison of decrements caused by variation in the irrelevant dimensions. The left panel varied the fundamental frequency as the irrelevant dimension (Experiment 2), whereas the right panel varied the number of spectral peaks as the irrelevant dimension (Experiment 3).


75
70
65
60
55
50
Intensity (dB)
0 2 4 6 8 10 12 14 16 18 20
Harmonic Numbers

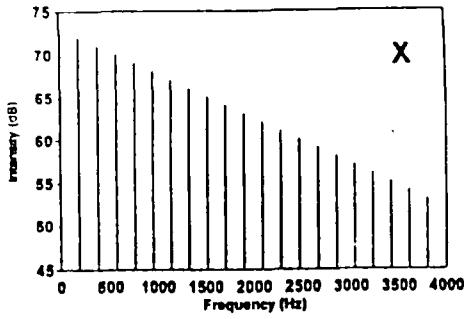
(a)


75
70
65
60
55
50
45
40
Intensity (dB)
0 2 4 6 8 10 12 14 16 18 20
Harmonic Numbers

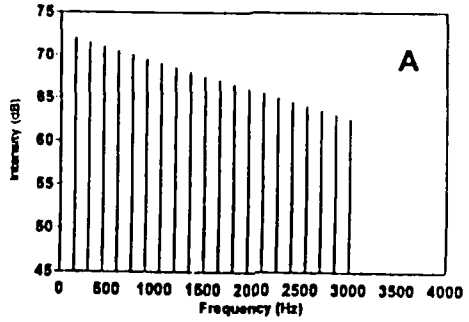
(b)

Fig. 1

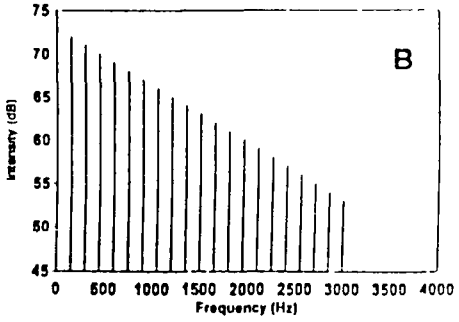
Irrelevant dimension = fundamental frequency



F0 = 190 Hz
Slope = -1.0

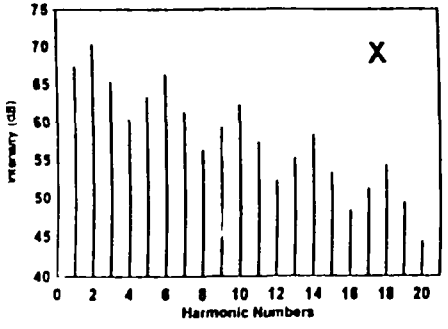


F0 = 170 Hz
Slope = -0.5

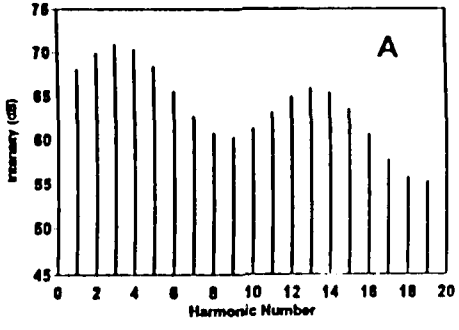


F0 = 170 Hz
Slope = -1.0

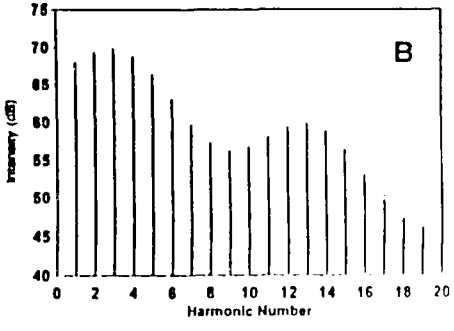
Irrelevant dimension = filter property (# of spectral peaks)



5 peaks
Slope = -1.0



2 peaks
Slope = -0.5



2 peaks
Slope = -1.0

Fig. 2

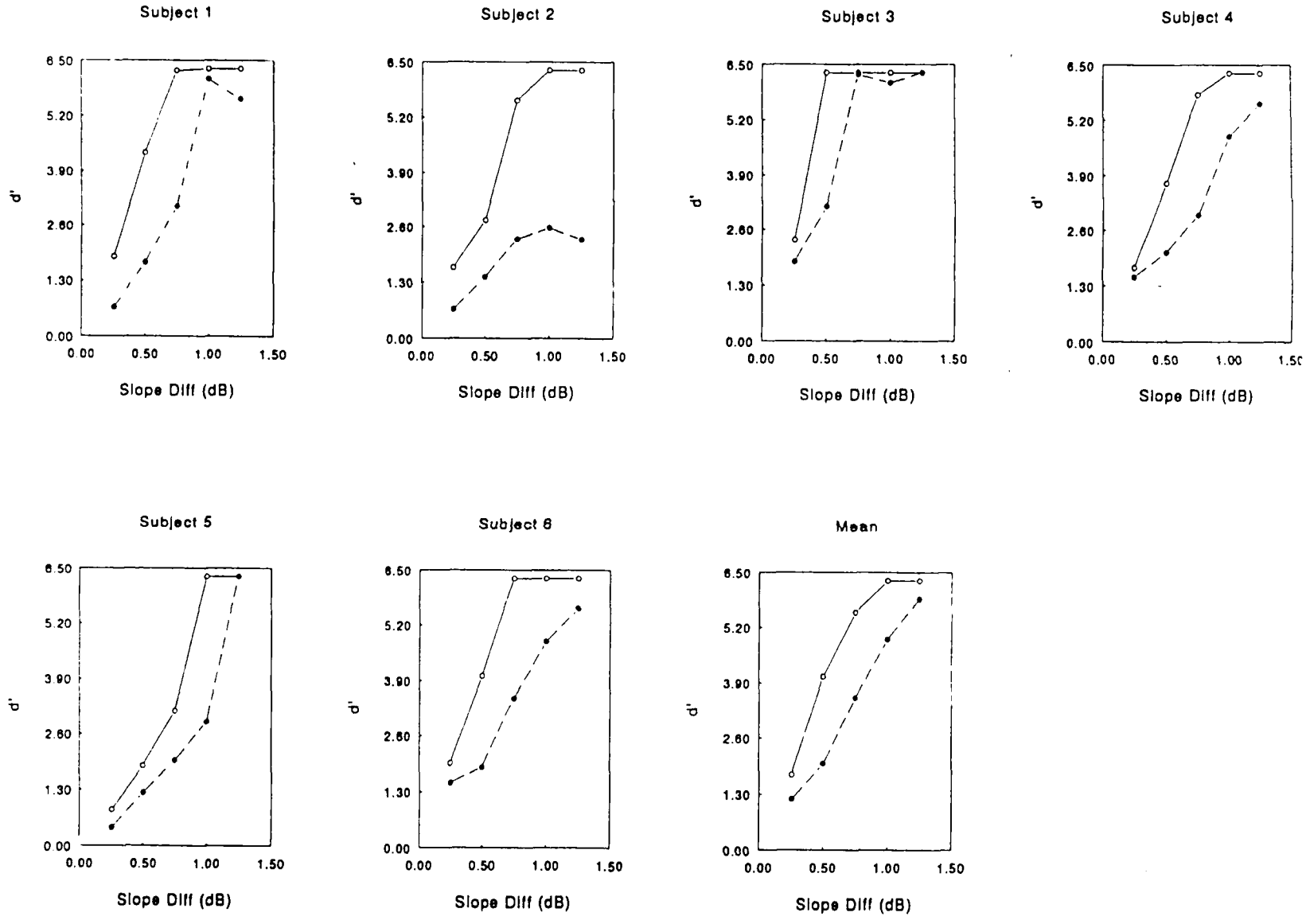

Subject 1
Subject 2
Subject 3
Subject 4
Subject 5
Subject 6
Mean
d'
6.50
5.20
3.90
2.60
1.30
0.00
0.00
0.50
1.00
1.50
Slope Diff (dB)

Fig 3

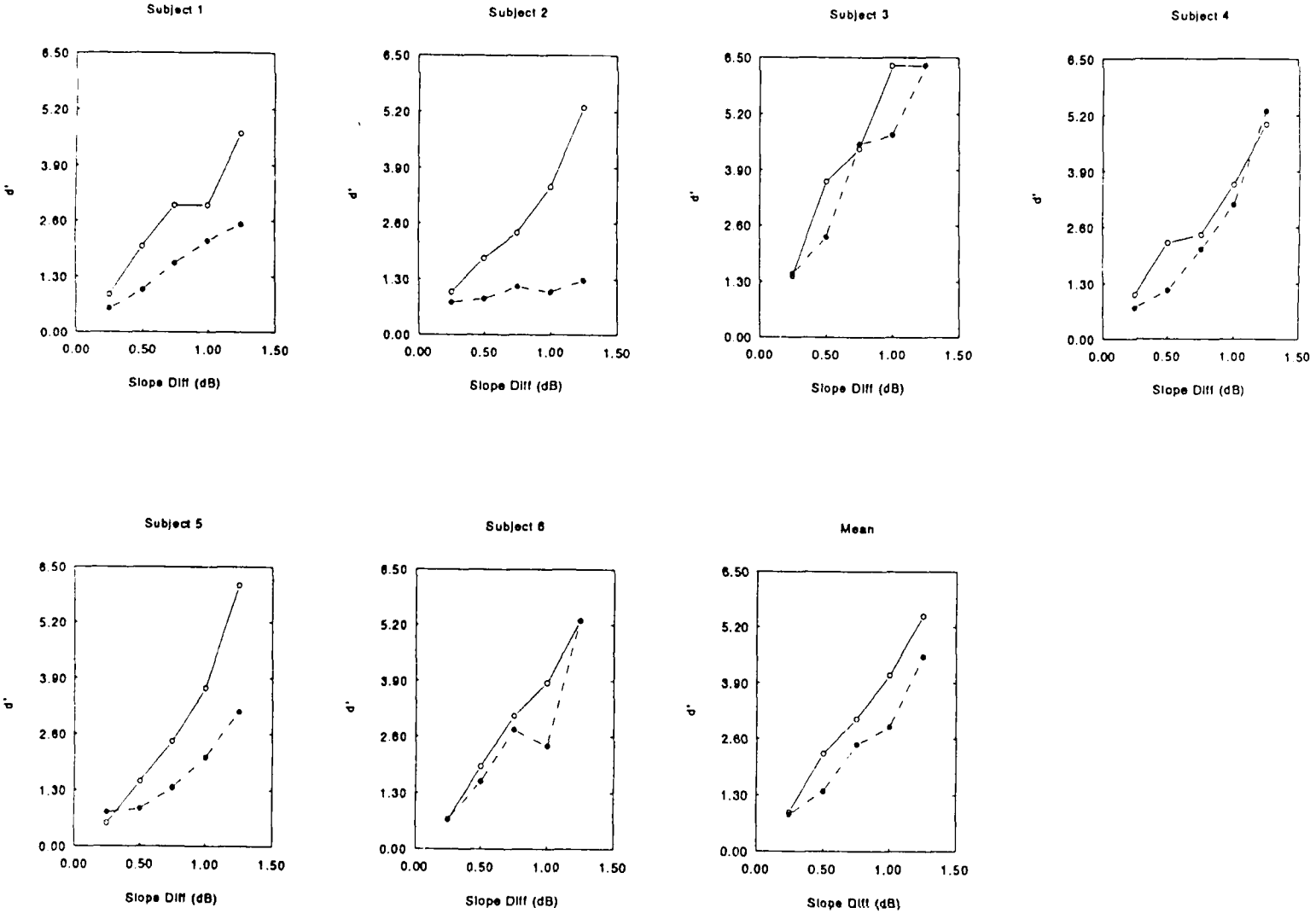

Subject 1
Subject 2
Subject 3
Subject 4
Subject 5
Subject 6
Mean
d'
6.50
5.20
3.90
2.60
1.30
0.00
0.00
0.50
1.00
1.50
Slope Diff (dB)

Fig 4

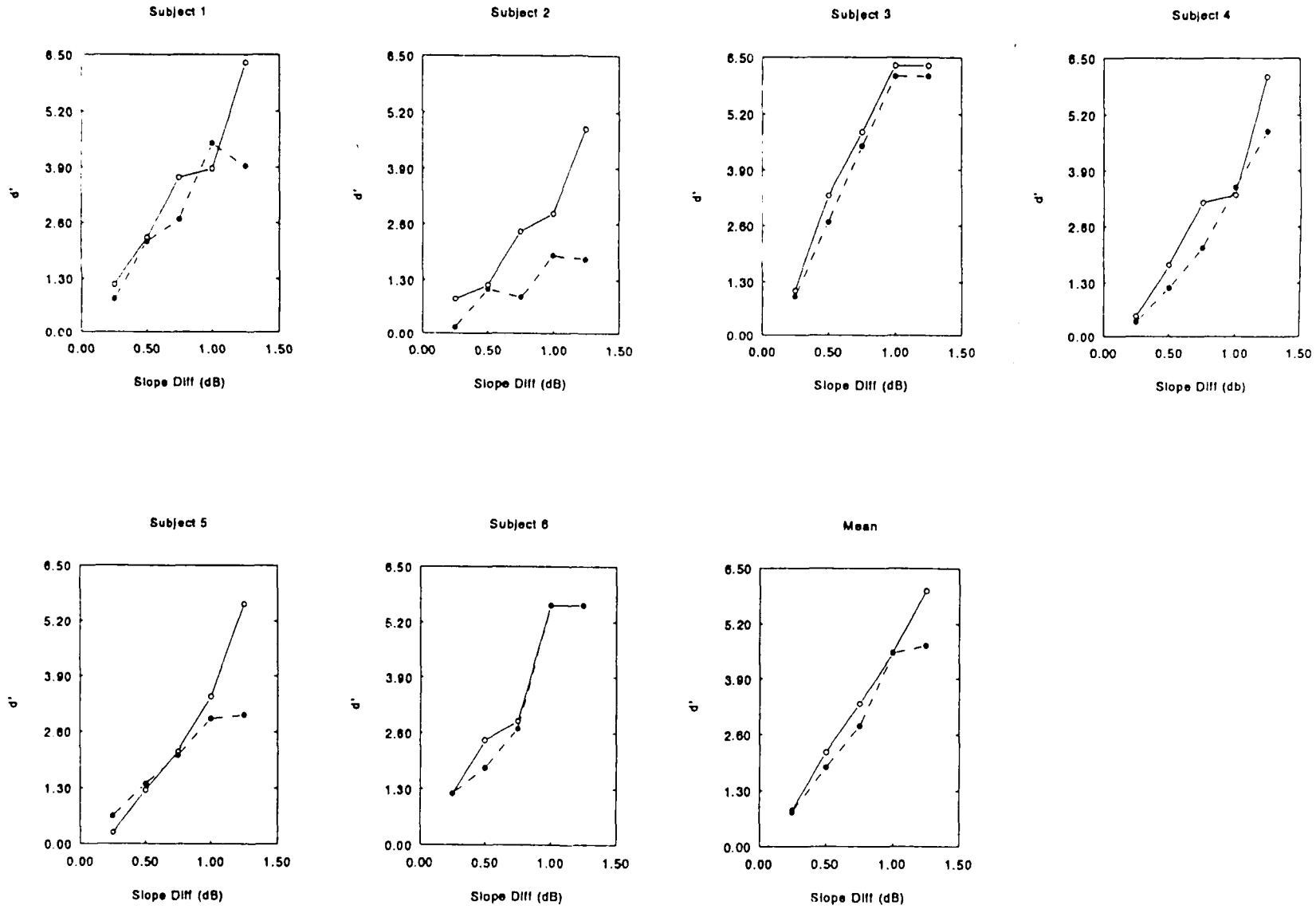

Subject 1
Subject 2
Subject 3
Subject 4
Subject 5
Subject 6
Mean
d'
6.50
5.20
3.90
2.60
1.30
0.00
0.00
0.50
1.00
1.50
Slope Diff (dB)
Slope Diff (db)

Fig. 5

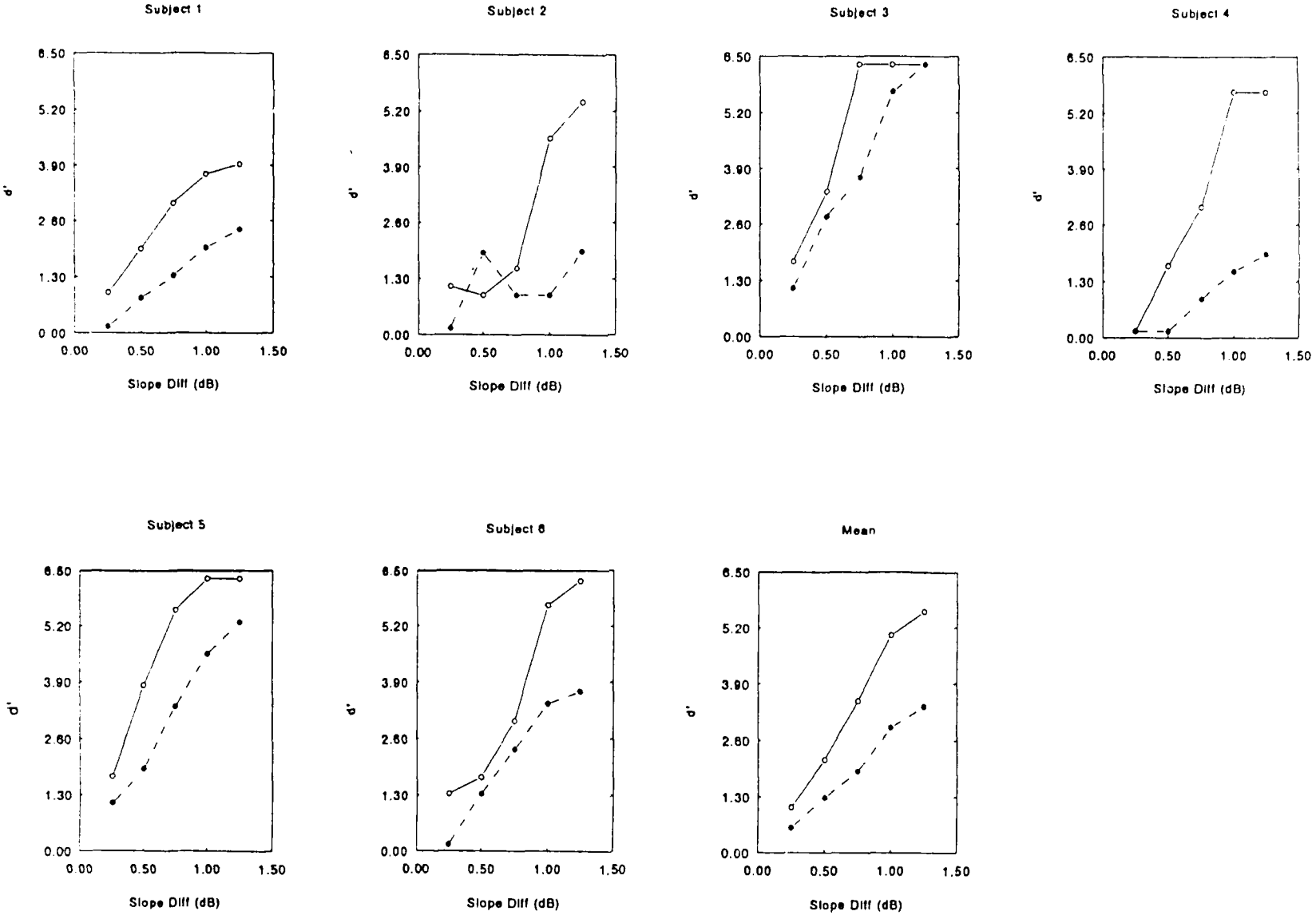
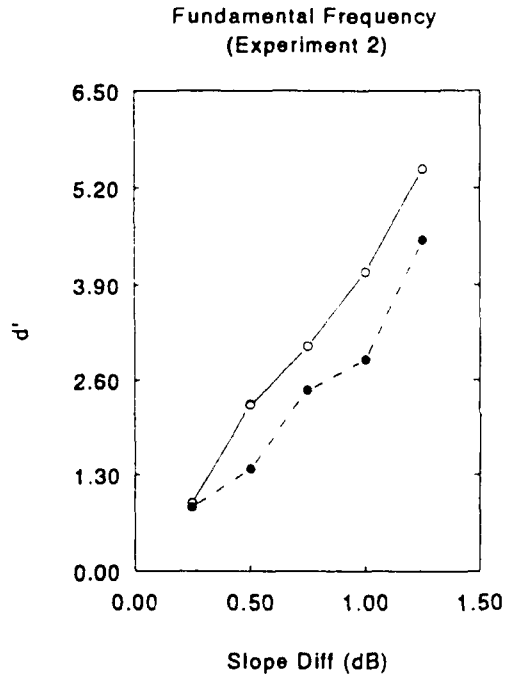

Subject 1
Subject 2
Subject 3
Subject 4
Subject 5
Subject 6
Mean
d'
6.50
5.20
3.90
2.60
1.30
0.00
0.00
0.50
1.00
1.50
Slope Diff (dB)

Fig 6


Fundamental Frequency
(Experiment 2)
6.50
5.20
3.90
2.60
1.30
0.00
d'
0.00
0.50
1.00
1.50
Slope Diff (dB)

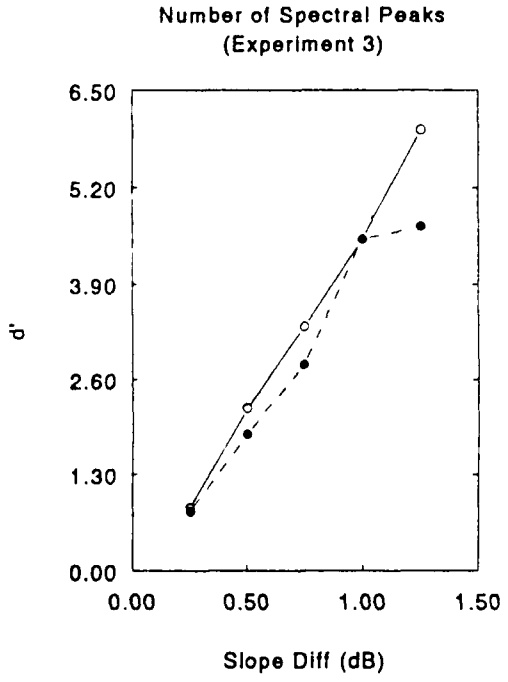

Number of Spectral Peaks
(Experiment 3)
6.50
5.20
3.90
2.60
1.30
0.00
d'
0.00
0.50
1.00
1.50
Slope Diff (dB)

Fig 7.

Within-category discrimination of musical chords: Perceptual magnet or anchor?

Barbara E. Acker, Richard F. Pastore, and Michael D. Hall

Abstract

Recent speech research (e.g. Lauckner-Morano & Sussman, 1993; Lively, 1993; Kuhl, 1991) has demonstrated that the presence of prototypes may be reflected in the internal structure of speech categories. The current study examines the function of prototypes in another natural, but nonspeech category. Prototype (P) and nonprototype (NP) sets of major triad stimuli were constructed, with stimuli in the P set being more representative of the category than the NP stimuli. Musically experienced subjects rated the stimuli in each set for goodness as a major triad, with the highest rated stimulus serving as a prototype standard for a subsequent discrimination task. Results from the discrimination task demonstrated better performance in the prototype context. Some contrasting speech results (Kuhl, 1991) instead found lower discrimination in a vowel P context compared to a NP context, suggesting that a prototype may function as a perceptual magnet, effectively decreasing perceptual distance, and thus, discriminability, between stimuli. The current nonspeech results appear to follow predictions based on classification and perceptual models, and provide a natural, nonspeech contrast to speech findings.

End of Abstract

Running Head: Chord Discrimination

Although it has long been conjectured that speech categorization may be based upon the use of prototypes or exemplars, most speech perception research has focused on the location of category boundaries, with little attention given to perceptual changes which, in theory, should exist within categories that are based upon the use of prototypes. This focus on category boundaries probably is a carry-over from the notions of categorical perception which posited absolute recoding of perception in terms of discrete phonetic categories, with any within category variation dismissed as stimulus artifacts (Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). In contrast to this long tradition of categorization studies based on labeling tasks, some recent research has begun to examine the internal structure of speech categories (Kuhl, 1991; Lauckner-Morano & Sussman, 1993; Li & Pastore, 1992; Lively,1993; Samuel, 1982; Volaitis & Miller, 1992). In general, category membership is found to be qualitatively graded, with Kuhl (1991) providing evidence that quality of membership is reflected in the discriminability between stimuli. Qualitative grading and patterns of discrimination within categories are probably indicative of general category structure and should also occur in nonspeech categories. The current research evaluates quality of category *membership* and discriminability for a musical category, another natural, but nonspeech category.

Prototype and Exemplar Notions

Prototype and exemplar models are currently the two more predominant approaches to modeling perceptual categorization, each addressing the nature of category members somewhat differently. Prototype theory (e.g. Posner & Keele, 1968) attributes categorization to the comparison of incoming stimuli to internal prototypes which are some form of averaged or ideal category representations. Although the exact nature of a prototype may differ across models, there is some agreement that experiences with a particular category contribute to defining the category prototype, and that categorization usually is assumed to be determined in terms of the relative match (or perceptual distance) between the incoming stimulus and the prototype. As similarity to the prototype decreases (i.e., as perceptual distance increases), the quality of category membership decreases, and the probability of assignment to the given category should decrease. Furthermore, as stimuli increase in similarity to (and thus, decrease in perceptual distance from) the prototypes for other categories, the probability of assignment to other categories should increase.

Exemplar theory (e.g. Medin & Schaffer,1978; Nosofsky, 1991) proposes that experiences are stored in memory and categorization is determined by the set of the exemplars elicited by the incoming stimulus. Exemplars for a given category are specific instances of a stimulus, rather than a single, averaged representation of experienced stimuli. In modeling categorization, the incoming stimulus is assumed to be categorized according to the degree of similarity to the stored exemplars. If similarity is high relative to the pool of exemplars for a given category, the incoming stimulus will be assigned to that category exemplar. Although it has often been argued that categorization findings are better described by exemplar, rather than prototype models, most such research has focused on limited, often artificially-defined categories. Li & Pastore (1992) note that differentiating between exemplar and prototype models for natural categories (e.g. stop consonants) can be very difficult (also see Nosofsky, Palmeri, & McKinley, 1993). Furthermore, for highly learned categories (such as pitch categories for musicians), assuming the storage of memories of all specific instances appears to require an unreasonable memory load and lengthy search processes, and is therefore difficult to model.

One reasonable approach to modeling exemplar notions about natural or highly learned categories would be to focus on the central tendencies in the probability distribution of sampling exemplars. It is logical to assume that categories should have a higher concentration of good exemplars than poor exemplars, and therefore, the probability of these good exemplars being elicited by the incoming stimulus should increase. Although the difference between exemplar and prototype notions is theoretically important for understanding categorization processes, this more global level of analysis in the current manuscript does not allow for differentiation between the two [illegible] the term prototype will be used throughout the current discussion [illegible]

evaluating stimuli by sampling from a pool of stored exemplars rather than by comparison to a prototype.

Several predictions can be made in regard to the pattern of discrimination within a category structured around a prototype. Viewed in terms of the auditory perception model proposed by Braida, Lim, Berliner, Durlach, Rabinowitz, & Parks (1984), a prototype should function as an interior anchor. Discrimination should be enhanced for a prototype set of stimuli (stimuli near a prototype). In contrast, a nonprototype (NP) set should consist of stimuli which, while from the same natural category, are distant from the given category prototype (and any alternative category prototypes) and are poor examples of the category. These stimuli should have been experienced less frequently, and thus should have a sparse distribution of similar exemplars which are not common representatives of the category. Therefore, there is no reason to posit the formation of anchors in a NP context. Thus, if anchors provide a basis for discrimination, the absence of an anchor in a NP context should result in poor discrimination between stimuli in a NP category.

Following another line of reasoning, if discrimination is modeled on the basis of perceptual distance from a standard (e.g. a prototype or anchor) and obeys any sort of approximation to Weber's law, stimuli differing by a constant amount (Δd) from each other should be most discriminable if they are similar to and thus, a small distance from the prototype (e.g. with a constant Δd, discrimination should be an inverse function of d, the distance from the prototype). Thus, stimuli around a prototype again would be predicted to be more discriminable than those sampled at some distance from a prototype.

The prototype (P) and nonprototype (NP) sets of stimuli in the current study are constructed with a number of common stimuli, which can be used to evaluate differences in the scale of goodness ratings for the P and NP sets of stimuli. (See the circles in Fig 1 for examples of common stimuli.) Some general ideas from the Macmillan, Braida, & Goldberg (1987) model of auditory perception, based upon trace and context coding concepts, would predict context differences in the goodness ratings of stimuli common to the P and NP sets. Goodness ratings should be based upon comparisons made with respect to the entire range of stimuli used in the experiment. Furthermore, stimuli in a P context probably should reflect considerable prior knowledge and learning because they include common, frequently experienced, representatives of the category. In contrast, stimuli from a NP context would not reflect prior knowledge and learning, as they are poor representatives of a category and are not often encountered. These experiential differences should be reflected in different ratings for any stimuli which are common to both the P and NP contexts.

It is known that ratings can be affected by the range and distribution of stimuli (Parducci & Perret, 1971). Although the two sets of stimuli in the current study are equal in the size of physical range, the distribution of perceived quality relative to a prototype anchor should not be equal. These differences in qualitative distributions should result in the differential use of rating scales according to the context which is being judged, reflected in the different ratings of stimuli common to the P and NP contexts. In particular, a stimulus rated relatively low in a P context should be rated higher if presented in a NP context, where it is relatively better than the other members of the NP stimulus set. In summary, predictions for goodness ratings in a category anticipate different ratings according to context, and several classification models predict increased discrimination in a prototype context.

The "Perceptual Magnet" Effect

Considering these predictions, results from recent speech research (Kuhl, 1991) are somewhat unexpected. In examining the internal structure of a vowel category for stimuli consisting of a continuum of /i/ vowels varying in F2 and F3, Kuhl (1991) demonstrated a "perceptual magnet" effect. Kuhl constructed two different sets of /i/ stimuli, with the set of stimuli in the (P) context being more representative of the /i/ category than the set of stimuli in the (NP) context. Subjects rated the stimuli in each context for goodness as an /i/ vowel, including a sampling of stimuli common to both types of contexts. The theoretical prototype, based upon prior research (Peterson & Barney, 1952), was the highest rated stimulus across both the P and NP contexts, and was used as the standard for a subsequent discrimination task in a prototype context. The discrimination standard for the NP context was a low rated stimulus. Discrimination of other stimuli from the given context was measured as a function of distance in perceptual space (defined in mel units) from the standard. For equal distances in mel units, discrimination was found to be lower for the prototype relative to the nonprototype standard. These results were interpreted as reflecting reduced discrimination in the region of the prototype, with the prototype acting as a type of perceptual magnet that reduces the perceived distance of neighboring stimuli.

Neither consistent goodness ratings of common stimuli across context nor reduced discrimination near a prototype are predicted from various categorization models (Braida et al.,1984, Parducci & Perret, 1971) or psychophysical laws (e.g. Weber's law) summarized above. As a result, several attempts have been made to replicate the perceptual magnet effect for vowels. Lively (1993) found that goodness ratings differed across both subjects and context. Additionally, in contrast to a magnet effect, Lively found slightly heightened discrimination around an "average" prototype standard. When individual prototypes were used as a standard, discrimination was equal across contexts. Lauckner-Morano & Sussman (1993) partially replicated the Kuhl study, but questioned labeling the results as a special effect: an additional identification task provided some evidence that Kuhl's nonprototype stimuli could have included two perceptual categories, with improved discrimination due to the measurement of between, not within, category discrimination.

Following the general logic and procedures of the previous speech studies, the current study attempts to evaluate qualitative [illegible] discrimination in a nonspeech category. Prior research has demonstrated that a number of speech phenomena, such as [illegible] (Burns & Ward, 1978; Locke & Kellar, 1973; Siegel & Siegel, 1977) [illegible] (Hill & [illegible] 1982) [illegible] Rosenbaum & Szczesiul, 1978) [illegible]

unique to speech, the results of the current study therefore are important in evaluating the general nature of categorization for a natural category, and in serving as a comparison for the speech work.

General Method

Subjects

An assumption of prototype (and exemplar) theory is that experience is needed for the formation of good category representations (or a reasonable pool of exemplars). If knowledge of a particular category is minimal, experience with the category is probably insufficient to have formed a prototype. Therefore, we used 5 musically trained subjects who had a minimum of 10 years of experience, with 2 subjects having college-level training. Subjects were paid $5/hour for their participation.

Because the concepts of prototypes and exemplars anticipate individual differences, our musical rating task maintained a within subjects design, with each subject providing separate ratings for P and NP sets of stimuli (Experiment 1), and then performing a discrimination task (Experiment 2). By using a within subjects design, the current study avoids potential individual difference problems present in a between subjects design, previously used in some speech research (Kuhl, 1991). For example, if ratings were averaged across subjects in a rating task to determine a prototype standard, the averaged prototype standard may not be representative of the prototype of an individual subject, and the subsequent discrimination results may not accurately reflect the individual's category structure relative to this prototype. In contrast, the use of individual prototype standards for a discrimination task should more accurately reflect individual category structure.

Stimuli

Two sets of stimuli were constructed by generating individual sine tones (12 bit, 10 kHz sample rate) and digitally mixing them to form triads based on a root position C major triad (see Figure 1). The initial (theoretical) prototype stimulus was a perfectly tuned C major triad (C = 262 Hz, E = 330 Hz, G = 392 Hz). The other stimuli were generated by holding the C constant and varying the E and G frequencies in both sharp and flat directions in 2 Hz increments. For 8 of the 30 stimuli, only the E frequency varied, for another 8 stimuli, only the G frequency varied. Positive and negative diagonals were created by varying the E and G frequencies simultaneously with equal steps in either the same (both flat or sharp) or opposite (one flat, the other sharp) directions. Thus, the prototype stimuli ranged from 322 Hz to 338 Hz for the E and 384 Hz to 400 Hz for the G. Nonprototype stimuli, which were created in a similar manner, were based on a mistuned C major triad (C=262 Hz, E=338 Hz, G=384 Hz) and ranged from 330 Hz to 346 Hz for the E, and 376 Hz to 392 Hz for the G. Using these procedures, 30 stimuli differing in equal steps were created for the prototype and the nonprototype space, as summarized in Figure 1. Of the 30 stimuli in each set, 7 occurred in both sets (see circles in Fig. 1). The sampling of stimuli generally followed that used by Kuhl (1991) for speech stimuli.

Because of the limited frequency range, equal changes in frequency provided a close approximation to equal changes in psychophysical distance. Therefore, the cents scale was not needed to equate perceptual distance. All stimuli were 1000 ms in duration, were low-pass filtered at 4kHz, and presented over TDH-49 earphones at 78 dB(A) in commercial sound chambers.

---

insert figure 1 here

Experiment 1

The goal of Experiment 1 was to use goodness ratings to evaluate qualitative grading in a musical category. In addition, each individual's highest rated stimulus was used to identify their prototype standard for a subsequent discrimination task (Experiment 2).

Procedure

Subjects were instructed to rate the goodness of each stimulus as a major chord on a scale of 1 (very poor) to 7 (very good). Ratings were indicated by button presses on a telephone keypad and were collected by computer. To become familiar with the stimulus range, subjects listened to the 30 stimuli from a given context once in random order without responding. All stimuli then were presented three more times to provide practice using the scale. Data finally were collected from 20 randomized repetitions of the stimulus set. The procedure then was repeated for the other context, with context order counterbalanced across subjects.

Results and Discussion

Goodness ratings for the musical task (a) declined systematically from the individually defined (i.e. highest rated stimulus) prototype and (b) improved when moving from the nonprototype center stimulus toward the prototype [for P and NP contexts respectively, $F(4,16) = 38.38$, $p < .0001$, $F(4,16) = 49.72$, $p < .0001$; see Figure 2, panels A and B]. As Table 1 shows, the highest rated stimulus varied across subjects, but always occurred within 1 or 2 city block steps (2 or 4 Hz) from the perfectly tuned triad. Ratings between the individual and theoretical prototype are significantly different [$F(1,4) = 22.17$, $p < .01$]. The absence of perfect tuning in each individual's prototype was not completely surprising, as several studies have shown that musical intervals are often compressed (Rakowski, 1976) or stretched (Hartmann, 1993; Ward, 1954). Speech research (Lively, 1993) also found ratings to differ across subjects.

The rating of the theoretical prototype also significantly differed across contexts [$F(1,4) = 35.61$, $p = .0035$]. In the nonprototype context, the fact that the perfectly tuned triad received the highest average goodness rating is consistent with it clearly being the best exemplar available. The ratings of the other 5 shared stimuli also significantly differed across contexts [$F(5,20) = 11.44$, $p < .0001$], with all shared stimuli [illegible] rated higher in the NP context [$F(1,4) = 11.45$, $p < .05$]. [illegible] the nonprototype set was

also significantly different across contexts [$F(1,4) = 33.08$, $p < .005$], but overall, received lower ratings than the theoretical prototype. Recent speech research (Lively, 1993) also found that vowel goodness ratings of shared stimuli differed across context.

Insert Table 1 here

Because subjects differed in the average and range of rating employed in the P and NP contexts (reflecting differences in the use of the scale), it was difficult to legitimately either pool results across subjects or to compare results across conditions (P or NP contexts). In order to better equate the ratings across subjects, individual ratings for the P context were normalized to yield a mean rating of 3.5 (center of rating range) and a standard deviation of 1.8 (selected to keep all ratings between 1 and 7). These normalized data then were averaged to yield the rating results shown in Figure 2. The individual data for the NP context also were normalized for a mean of 2.03, thus matching the rating of the theoretical prototype across the two contexts) and a standard deviation of 1.09 (the original SD average across all subjects in the NP context). While the normalized results do not accurately reflect obtained rating differences between the theoretical prototype and nonprototype standards, they do reflect the relative rating tendencies both within and across contexts, with stimuli in the NP context being rated much lower than those in the P context.

Insert Figure 2

Summary. Individual differences in optimum stimulus ratings were expected and can probably be attributed to experiential factors, such as differences in the major instrument of study for the 5 subjects. In addition, most individual prototypes were slightly flat compared to the theoretical prototype. This slightly flat mistuning is consistent with the tradition of equal temperment, which slightly compresses major thirds and slightly stretches major fifths. Rating differences for common stimuli across contexts are consistent with predictions based on the range and qualitative distribution of stimuli.

Experiment 2

Experiment 1 demonstrated that musical categories are clearly qualitatively graded, with one stimulus always receiving a distinctively higher rating than the other category members. (The specific stimulus optimum stimulus may vary by individual.) Furthermore, the goodness ratings of stimuli decrease in a relatively systematic fashion as a function of distance from the prototype. It can be inferred from these findings that musical categories may be structured around prototypes or *frequent exemplars* (*i.e., the highest rated* stimulus). Therefore, a discrimination task similar to that used in several speech studies (Kuhl, 1991; Lively, 1993) should be able to effectively evaluate the function of a prototype in this natural, nonspeech category. Classification models (Braida et al., 1984) and Weber's law predict that discrimination should be higher in a prototype context than in a nonprototype context.

Procedure

Because the musical rating task revealed individual differences, we used a within-subjects design for a discrimination task, with each subject's highest rated stimulus from the goodness rating task employed as that subject's prototype standard. Additional support for using a within subjects design comes from Lively (1993), where the pattern of discrimination differed slightly, depending on whether an averaged prototype or the individual's prototype was used as a standard. Since for most of our subjects, the prototype was within 1 city block step (2 Hz) away from the perfectly tuned triad (which was the center of the P stimulus set), the nonprototype standard was selected to be one step from the center of the NP stimulus set. Unlike the prototype standard, the nonprototype standard has no special significance other than representing a control condition, so there was no need to modify the selected standard for individual subjects. Therefore, a single nonprototype standard was used for all subjects.

Because Kuhl (1991) wanted to use a single procedure for human infants and adults, as well as monkeys, she used a go, no-go procedure. However, this procedure is highly sensitive to response bias. A go, no-go task is essentially equivalent to an ABX task, and a recent study of ABX discrimination of stimuli drawn from a continuum between major and its related minor triads (Howard, Rosen, & Broad, 1992) demonstrated significant bias. Therefore, we used a 2IFC task, which should provide a more reliable discrimination measure for each stimulus (28 repetitions for each stimulus vs. 2 in the Kuhl (1991) discrimination task) and which is bias-free.

Subjects heard two pairs of triads per trial, consisting of a same and a different pair. In each pair, the standard always was presented first. In "same" pairs the standard (P or NP) was repeated. The second stimulus in the "different" pair was randomly selected from the 29 other triads. The ordering of pairs within a trial was randomized. Subjects indicated with a button press which pair contained the different triads. The experiment was run in two sessions, with each session containing both the P and NP contexts. Selection of initial context (P or NP) was counterbalanced across subjects, and the context initially presented in the first session was presented last in the second session. Subjects were presented each triad a total of 28 times.

Results and Discussion

Discrimination results for both contexts are shown in Figure 3. As predicted by traditional notions of anchors and perceptual distance, discrimination was higher in the P context than in the NP context. Average discrimination at one city block step from the major triad [illegible] when measured in a positive (upward, sharp) direction and at [illegible] when measured in a negative (downward, flat) direction. [illegible] individual differences [illegible] level of performance [illegible]

similar trend, with performance jumping to near perfect performance for a difference of two city block steps (4 Hz change) from the standard. In fact, one subject demonstrated virtually perfect discrimination in the P, but not NP context. All of these findings are consistent with the predictions from categorization and perceptual models.

Insert Figure 3 here

General Discussion

The current study investigated aspects of perception for stimuli varying systematically in the frequency of two steady-state components. Goodness ratings revealed that the theoretical musical prototype (the perfectly tuned triad) was never rated the highest in the P context and that the optimum stimulus varied across subjects, but that all stimuli common to both sets (P and NP) were rated higher in the NP context. As discussed in the introduction, context effects such as these are consistent with at least two similar concepts. First, ratings have been shown to be affected by the range and distribution of stimuli. While the music stimuli contexts (P or NP) were of equal physical range, qualitative perceptual range was not equal. Therefore, different ratings for the same stimuli embedded in different contexts are expected. Similarly, different ratings for common stimuli across different contexts (P or NP) are consistent with context coding concepts, where comparisons of stimuli are made with respect to the entire range of stimuli used in the experiment, and thus should be a function of stimulus set.

Discrimination for equal distances from a standard in a musical category was better in a prototype context than in a nonprototype context. While in contrast to predictions based on a perceptual magnet, these discrimination findings from a natural category are consistent with typical psychophysical laws and confirm discrimination is probably achieved with the use of some form of anchors or reference points (Braida et al., 1984; Macmillan, Braida, & Goldberg, 1987).

Comparison to Recent Speech Studies

Recent speech research investigating qualitative grading in a vowel category (Kuhl, 1993; Lively, 1993) has shown ratings of optimum stimuli to differ across individuals. The current findings for a musical category, along with these recent speech results, indicate that some individual differences are probably typical of most general perceptual processes, and future research should investigate the small discrepancies between true individual prototypes and theoretical prototypes.

Lively (1993) found goodness ratings of vowel stimuli common to both contexts (P and NP) to be all rated higher in the NP context than the P context. In contrast, the original vowel rating study (Kuhl, 1991) did not demonstrate clear rating differences across contexts, and the same stimulus received the highest rating in both contexts. Variable ratings across contexts are consistent with predictions based on the range and qualitative distribution of stimuli, while consistent ratings across contexts are not.

The differences across the speech and nonspeech studies found in the goodness ratings may be a function of stimulus selection. While the present study shows a consistent pattern of ratings for stimuli common to both sets, it cannot directly evaluate the stability of ratings for the stimuli which served as the actual prototype stimulus for the various subjects. Due to the construction of the prototype stimulus set based upon the individual prototypes, which were never the theoretical prototype, the actual individual prototype was never present in the NP context. If each individual's prototype had been present in the NP context, it is possible that this stimulus would have functioned as a perceptual anchor, thus removing range effects.

Discrimination results from the current study differ from the summarized speech findings. Kuhl (1991) found better discrimination in a NP context than a P context, and interpreted these results as demonstrating reduced performance in the P context, with the prototype exerting a magnet effect on surrounding stimuli, as opposed to enhanced performance in a NP context. While the Kuhl conclusion is not consistent with either a Weber function or anchor and reference point notions, it seems to be based upon the logical assumption that, other factors being equal, performance differences should be attributed to the action of a prototype, rather than some unknown factor enhancing performance in the NP context. This logical reasoning leads to the conjectured "perceptual magnet" metaphor. In an attempted replication, Lively (1993) found slightly better discrimination in a P context with the prototype standard defined by Kuhl, and when individual prototypes were used as a standard, discrimination was virtually equal across contexts. Neither discrimination results are consistent with classification and perceptual models, which predict heightened discrimination in a P context, as found in the current study.

Why do the music results differ from the original speech results in terms of a perceptual magnet effect for the prototype? One possibility is that music and speech categorization processes are qualitatively different. Alternatively, it may be possible that Kuhl (1991) had actually found enhanced discrimination in a NP context, not reduced discrimination in a P context. As noted above, Lauckner-Morano & Sussman (1993) used a labeling task to demonstrate that Kuhl's original speech nonprototype context may have contained two categories. If this indication is an accurate assessment of the original NP context vowel stimuli, then Kuhl's original findings may not be indicative of a magnet effect, but rather of artificially enhanced discrimination in the nonprototype context due to between category comparisons. Thus, additional speech research is necessary to study the nature of prototype and nonprototype discrimination under acceptably equivalent conditions. The present results from another natural category, while not addressing the validity of the speech findings, support the existence of a prototype effect in the category structure for musical triads and can also serve as a comparison for future speech findings.

References

Braida,L.D., Durlach, N.I., Lim, J.S., Berliner, J.E., Rabinowitz, W.M., & Purks, S.R. (1984). Intensity Perception. XIII. Perceptual anchor model of context coding. *Journal of the Acoustical Society of America*, **76**, 722-731.

Burns, E.M., & Ward, W.I. (1978). Categorical Perception - phenomenon or epiphenomenon. Evidence from experiments in the perception of melodic musical intervals. *Journal of the Acoustical Society of America*, **63**, 456-468.

Grieser, D.L. & Kuhl, P.K. (1989). Categorization of speech by infants: Support for speech-sound prototypes. *Developmental Psychology*, **25**, 577-588.

Hall, M.D. & Pastore, R.E. (1992). Musical duplex perception : Perception of figurally good chords with subliminal distinguishing tones. *Journal of Experimental Psychology: Human Perception and Performance*, **18**, 752-762.

Hartmann, W.M. (1993). On the origin of the enlarged melodic octave. *Journal of the Acoustical Society of America*, **93**, 3400-3409.

Kuhl, P.K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototype of speech categories, monkeys do not. *Perception and Psychophysics*, **50**, 93-107.

Davis, K. & Kuhl, P. (1993, May). Acoustic correlates of phonetic prototypes: Velar stops. Poster presented at the 125th meeting of the Acoustical Society of America. Ottawa, Canada.

Lauckner-Morano , V., & Sussman, J.E. (1993, May). Identification and change/no-change discrimination of /i/ stimuli: Further tests of the "magnet" effect. Paper presented at the 125th meeting of the Acoustical Society of America. Ottawa, Canada.

Li, X. & Pastore, R.E. (1992). Evaluation of prototypes in perceptual space for a place contrast. In M.E.H. Schouten (Ed.). *The Auditory Processing of Speech*. New York: de Gruyter, 303-308.

Lively, S.E. (1993, May). An examination of the perceptual magnet effect. Paper presented at the 125th meeting of the Acoustical Society of America. Ottawa, Canada.

Locke, S., & Kellar, L. (1973). Categorical perception in a nonlinguistic mode. *Cortex*, **9**, 355-369.

Macmillan, N.A., Braida, L.D., Goldberg, R.F. (1987). Central and peripheral processing in the perception of speech and nonspeech sounds. In M.E.H. Schouten (Ed). *The Psychophysics of Speech Perception*. Dordecht, The Netherlands: Martinus Nijhoff Publishers.

Medin, D.L. & Schaffer, M.M. (1978). Context theory of classification learning. *Psychological Review*, **97**, 225-252.

Nosofsky, R.M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. *Journal of Experimental Psychology: Human Perception and Performance*. **17**, 3-27.

Nosofsky, R.M., Palmeri, T.P., McKinley, S. (1993, November). Rule-Plus-Exception model of classification learning. Paper presented at the 34th meeting of the Psychonomic Society, Washington D.C.

Parducci, A., & Perrett, L.F. (1971). Category rating scales: Effects of relative spacing and frequency. *Journal of Experimental Psychology Monographs*, **89**, 427-452.

Pastore, R.E., Schmuckler, M.A., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception with musical stimuli. *Perception and Psychophysics*, **33**, 469-474.

Peterson, G.E., & Barney, H.L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, **24**, 175-184.

Posner, M.I., & Keele, S.W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.

Rakowski, A. (1976). Tuning of isolated musical intervals. *Journal of the Acoustical Society of America*, **59**, S50 (A).

Samuel, A.G. (1982). Phonetic prototypes. *Perception and Psychophysics*, **31**, 307-314.

Siegal,, J.A., & Siegal, W. (1977). Categorical perception of tonal intervals: Musicians can't tell sharp from flat. *Perception and Psychophysics*, **21**, 399-407.

Studdert-Kennedy, M. Liberman, A.M., Harris, K.S., & Cooper, F.S. (1970). Motor theory of speech perception: A reply to Lane's critical review. *Psychological Review*, **77**, 234-249.

Volaitis, L.E. & Miller, J.L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America*, **92**, 723-735.

[illegible] Subjective [illegible] **26**, 369-384.

Table 1

Individual and Theoretical Prototype Ratings in a Prototype Context

	Subject 1	Subject 2	Subject 3	Subject 4	Subject 5
Individual Prototype Stimulus					
E Frequency	330 Hz	328 Hz	330 Hz	328 Hz	328 Hz
G Frequency	396 Hz	390 Hz	394 Hz	392 Hz	394 Hz
rating	**5.3**	**5.9**	**6**	**4.9**	**5.9**
Theoretical Prototype (E = 330 Hz, G = 392 Hz)	**4.3**	**5.6**	**5.1**	**4.1**	**4.5**

Acknowledgements

This research was supported by grant F496209310033 and F49069310327 from the Air Force Office of Scientific Research. The opinions expressed are those of the authors and do not necessarily represent those of the granting agency.

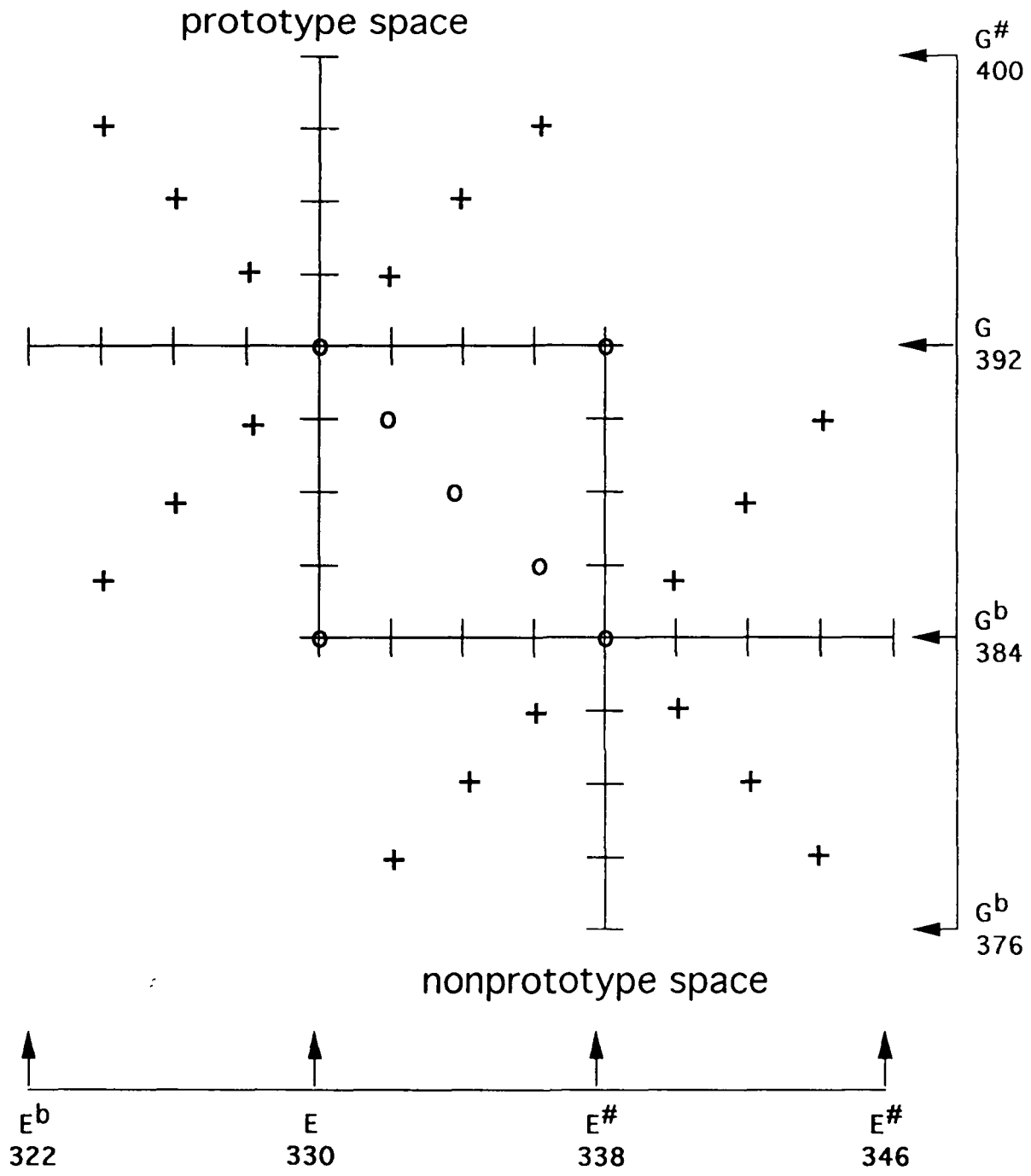
Figure Captions

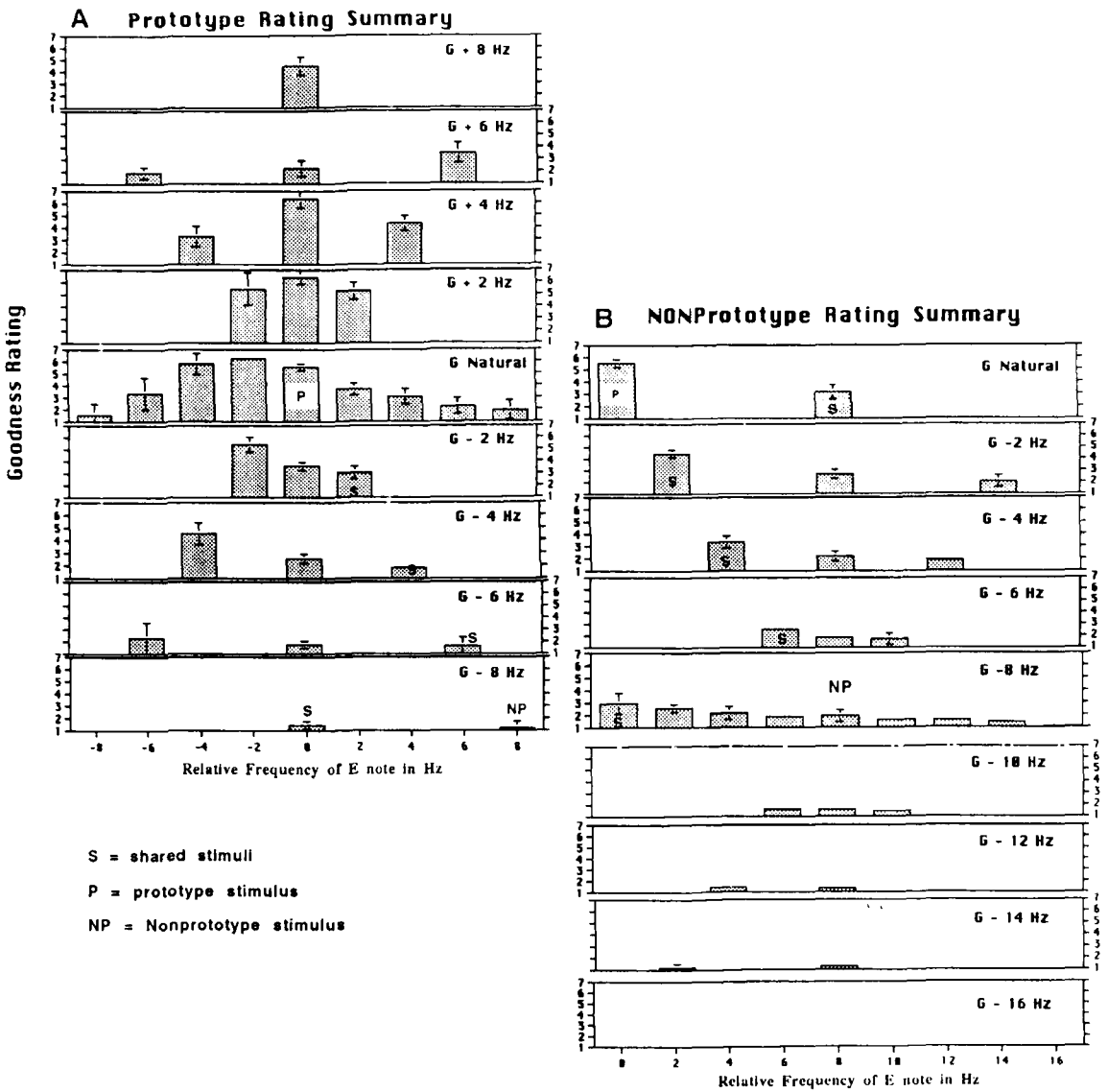
Figure 1: Perceptual space for C-major triads.

Figure 2: Normalized goodness rating summary for the prototype and nonprototype contexts. This figure parallels the structure of figure 1, and the shared stimuli are equivalent to the stimuli represented by circles in figure 1. Bars representing a rating of 1 do not appear.

Figure 3: Discrimination as a function of distance from a standard. Because individual prototypes were used, the same stimuli do not necessarily represent the same distances from individual prototypes. Therefore, some of the data points for the prototype standard may not include all five subjects.

Perceptual Space for C-major Chords




A Prototype Rating Summary
G + 8 Hz
G + 6 Hz
G + 4 Hz
G + 2 Hz
G Natural
G - 2 Hz
G - 4 Hz
G - 6 Hz
G - 8 Hz
Goodness Rating
Relative Frequency of E note in Hz
S = shared stimuli
P = prototype stimulus
NP = Nonprototype stimulus
B NONPrototype Rating Summary
G Natural
G -2 Hz
G - 4 Hz
G - 6 Hz
G -8 Hz
G - 10 Hz
G - 12 Hz
G - 14 Hz
G - 16 Hz
Relative Frequency of E note in Hz

Normalization of Musical Instrument Timbre

Jennifer L. Cho, Michael D. Hall, & Richard E. Pastore
Center for Cognitive & Psycholinguistic Sciences
State University of New York
Binghamton, NY 13902-6000

Abstract

The perceptual system appears to engage in an active, time-consuming process called normalization that maintains perceptual constancy by adjusting for source differences. Experiment 1 attempted to demonstrate music normalization for task-irrelevant timbre variability in a chord discrimination task. Experiment 2 demonstrated that music normalization is an active process. Experiment 3 identified important global components of instrument timbre which should be differentially subject to normalization. Based upon assumptions about the nature of speaker normalization, it was expected that perceived timbral similarity would produce faster response times, while dissimilar timbres would result in longer response times. A similarity scaling procedure was used to assess contributions of temporal (or attack) and spectral (upper harmonics) components to timbre for intact and physically-altered natural stimuli. Results indicate that, for the relatively long stimuli used in the current study, timbre was based primarily on the nature of the upper harmonics, with little contribution from attack functions. The relevance of these stimulus properties to normalization was evaluated in Experiment 4 using an AX chord discrimination task for a selected subset of Experiment 3 stimuli. Normalization, as indicated by RT, was inversely related to similarity. Information present in the higher harmonics also appeared to be most relevant to normalization.

Normalization is a type of perceptual constancy that can be loosely defined as the process by which the perceptual system adjusts for differences between sources in order to preserve an intended perceptual message. Normalization is an important concept in the speech perception literature where variability in vocal apparatus, context, and articulation are among the many relevant factors that determine the unique characteristics of words spoken by different talkers (e.g., Johnson, 1988; Jusczyk, Pisoni, & Mullennix, 1989; Mullennix & Pisoni, 1989; Mullennix, Pisoni, & Martin, 1989; Summerfield & Haggard, 1975). Despite extensive variability among speakers, listeners typically readily comprehend utterances produced by a wide range of talkers under a variety of conditions. This ability to recognize and adjust to task-irrelevant differences among speakers has led investigators to assume that there exists a mechanism that "normalizes" the disparities in speaker voice characteristics to efficiently maintain perceptual constancy in perceiving speech signals (e.g., Logan, 1989; Nusbaum & Morin, 1989).

Research has demonstrated that the normalization process is time-consuming and resource demanding. For example, Allard (1976), measuring reaction time (RT) in an AX task for word stimuli varying in speaker, noted that "same word" decisions tended to be faster when the two stimuli were physically identical relative to trials with changes in either speaker or intonation. Verbrugge et al. (1976) showed that identification of natural vowels was more accurate when the stimuli were tokens produced by a single talker rather than tokens produced by a number of talkers.

8 Recent concerns expressed by Goldinger (1992) have suggested that extant normalization research and theory have not created compelling arguments for a speaker-normalization process. The argument that effects of speaker variability are attention-demanding implies that speaker variability should impair performance of subjects operating under time constraints, even if no normalization process is engaged. According to Goldinger, virtually all "normalization effects" could be due to mere distraction from irrelevant source variability, rather than an actual normalization for speaker. In essence, simply demonstrating effects of stimulus variability does not distinguish between normalization as a separate process and normalization as a reflection of typical limitations on processing imposed by added stimulus variability. We use the term "passive normalization" for the latter conceptualization since the perceptual system is simply conjectured to respond directly to information in the stimulus.

The passive normalization hypothesis assumes that the perceptual system evaluates stimulus differences based upon some decision metric which is monotonically related to signal-to-noise ratio (S/N). Following a Signal Detection Theory analysis, the difference between any pair of stimuli defines the difference in central tendency between the distribution of same and different events and thus the magnitude of S. The decision system has noise due to the variability in the stimuli and stimulus coding, thus defining N. Adding variability to a given comparison (with fixed or constant S) thus increases N and results in a corresponding decrease in S/N. Slower and poorer performance would be due to the more difficult discrimination (reflecting the lower S/N). Therefore, according to this description, normalization is not a separate perceptual process. Rather, the decrements in performance simply reflect different levels of signal quality relative to noise. This conceptualization of normalization is theoretically uninteresting (as Goldinger might agree) because it implies that stimulus variability effects are not the result of some form of adaptive central processing, but instead reflect the standard operation of relatively static signal processing mechanisms.

An active normalization hypothesis begins with the conditions described for passive normalization, but assumes that the perceptual system is able to respond to, and then factor out, some expected task-irrelevant stimulus variability. The system is conjectured to accurately anticipate the nature of the variability, then effectively reduce N, thus increasing S/N. Incorrect anticipation can result in an inappropriate reduction of S, increase in N, or both, any of which may culminate in an overall decrease in S/N. Altering the system to accommodate the appropriate nature of the variability is assumed to require time and utilize processing capacity. The increase in RT would reflect the time needed by the processes to restore S/N, which, in turn, should result in high response accuracy. Consistent with this S/N characterization, Summerfield and Haggard (1975) have

suggested that reaction time results may reflect the natural operation of the normalization process, whereas the decrease in accuracy when responding to tokens produced by different sources may simply indicate that the normalization process may not be perfect. Thus, the accuracy effects observed in the speech literature may reflect some degree of a failure of normalize. In order to provide a convincing argument for any normalization process, one must demonstrate that the listener actively uses knowledge or memory of particular sources in adjusting for their variability.

Normalization and Instrument Timbre

The present investigation attempts to demonstrate the existence of an active normalization process in perceiving music stimuli. The use of music stimuli may offer an alternative forum in which to investigate the nature of normalization processes. Also, the sometimes simpler structure of music stimuli (Handel, 1989) may enable us to gain a better understanding of the processes implicated in normalization.

In speech normalization, variations among speakers are normalized in the perception of words; if there is an analogous music normalization process, then variations in instrument timbre should be normalized in the perception of triads. Timbre is the subjective attribute of source (instrument) that is based on invariant properties that uniquely characterize the tones produced by the source. Unfortunately, the pursuit of an adequate definition of timbre is both related to and dependent upon establishing which characteristics (or combination of characteristics) are important for perceptually determining an instrument's distinctive sound quality. As a result, existing "definitions" of timbre tend to focus as much upon what does not constitute timbre as what factors actually contribute to timbre. Thus, the American Standards Association (1960, p.45) defines timbre only as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness are dissimilar."

In addition to determining the possible nature of normalization, the current research also seeks to identify some important global components of instrument timbre that may be differentially subject to normalization. Speech normalization research to date has established that the auditory system adjusts for variability in the articulator (e.g., a speaker's vocal tract) (Nusbaum & Morin, 1989). However, only limited research has been conducted to specify what kind of variability the system is anticipating. By determining the bases of listener expectations and their relations to physical characteristics of the signal, we may also utilize normalization as a tool for investigating perceptual processing. In the process, we also should be able to effectively address the issue raised by Goldinger about whether the loss in speed is simply the result of (passive) perceptual limitations, or some active (normalization) process in which the system adjusts itself to factor out certain expected irrelevant properties.

Relevant Properties of Music Stimuli

An implicit assumption in most of the speaker normalization research has been that listeners make some judgment or comparison based on the similarity between two tokens (Logan, 1990). If the two tokens are highly similar, they should be judged as originating from the same speaker and would not be subject to normalization processes. If the tokens are highly dissimilar, they should be judged as originating from different speakers and would be subject to normalization. Paralleling this logic, the same assumption should apply for the postulated music normalization, substituting instrument timbre for speakers. However, in order to test this assumption, it becomes important to determine what stimulus properties define instrument timbre for listeners.

One promising approach to understanding the physical correlates of timbre has come from research on the perception of systematically altered stimuli that has identified some attributes of waveforms that may be important in instrument recognition. For each partial of synthetic stimuli (modeled after natural stimuli), Grey (1977) identified an attack transient, and intermediate steady-state, and decay. Removal of initial 20-50 ms segments of his 250 ms stimuli resulted in a significant impairment in the ability to identify different instruments. Therefore, it would appear that onsets (or attacks) of musical tones may contain essential cues for identification and discrimination of instrument timbre. Grey's findings seem to confirm the suggestion by Saldhana and Corso (1964) that onset cues appear to be more significant than offset cues in discrimination tasks for music stimuli.

Multidimensional scaling (MDS) also has been an effective approach to understand the spatial representations of a listener's similarity and difference judgments among a given stimulus set (Grey & Moorer, 1977). In MDS, perceived similarities or differences are used to represent subjective distance, and then to create a cognitive "map" that attempts to describe the perceptual relationships among stimuli. Plomp (1976) used MDS to investigate the perception of steady-state portions of nine synthesized instruments playing the same note. Scaling of timbral similarity judgments was highly correlated with the pattern of the spectral (as opposed to temporal) envelope. Therefore, MDS research indicates that the physical pattern of energy in the spectrum seems to form a basis of timbre judgments. Generally, similarity clustering seemed to correspond to class membership of the instruments (e.g., strings, brass, woodwinds), indicating that timbre is largely determined by spectral composition.

In order to evaluate the importance of onset transitions and upper harmonics (thus temporal and spectral envelope) in instrument timbre, the present study included an investigation of normalization for stimuli that have been physically altered. If normalization processes exist for music stimuli, the prime factors contributing to instrument timbre should be differences in attacks and/or spectral composition among instruments. These considerations guided our selection of instruments for the present investigation, and are the bases for the physical manipulations in later experiments.

Current Investigation

The current investigation was composed of three chord (triad) judgment experiments and one similarity scaling experiment. Experiment 1 was a chord judgment study (patterned after speech normalization studies) that attempted to demonstrate normalization effects for synthetic music stimuli. Experiment 2 was a chord identification study that used a subset of the same stimuli to evaluate whether normalization is an active process.

Experiment 3 was composed of a series of scaling conditions in which natural instrument tokens were used. In one

condition, natural and synthetic tokens were compared to assess generalizability of the results from Experiments 1 and 2 (where synthetic instruments were used). Two other scaling tasks were used to identify the importance of attack and spectral composition in defining instrument timbre for our stimuli. Experiment 4 was another chord judgment experiment where a selected subset of the natural and altered-natural stimuli were used to evaluate the predicted relationships between perceived similarity and normalization. The information obtained from these studies begins to provide a better understanding of not only the nature of normalization, but also of which stimulus properties are "normalized."

Experiment 1

The goal of Experiment 1 was to demonstrate normalization for music timbre, evaluating whether changes in instrument alter the speed (but not the ability) of subjects to respond to the equivalence of triads in an AX task. If findings that characterize talker normalization also apply to music stimuli, judgment of triads should be faster when chords are played by the same instrument. Conversely, slower reaction times should be obtained when triads are played by instruments with significant timbral differences. Maintenance of performance accuracy in a multiple-source condition is critical to demonstrating the operation of normalization processes for the reasons described above for the active normalization hypothesis.

Method

Subjects. Twenty-five undergraduate students enrolled in psychology courses at Binghamton University participated in partial fulfillment of course requirements. Since subjects needed to distinguish between major and minor chords, all subjects were asked to self-select based upon having at least minimal knowledge of music theory. Because of past experience with the high variability in both the ability and motivation of subjects in this pool, we always establish a priori criteria for inclusion of subject data. For the current study, the criteria was better than chance performance in all the conditions. Four subjects failed to meet this criterion. Data from one other subject was lost due to a computer malfunction. Thus, results for this experiment are based on data from the remaining 20 subjects.

Stimuli. The stimuli consisted of chords by five digitally sampled instrument sounds produced on a Roland synthesizer keyboard. The five instruments were piano, harpsichord, violin, flute, and trumpet. These instruments were chosen on the basis of the characteristic physical properties of the stimuli they typically produce. Previous research seemed to identify the attack transition as an important property in defining instrument timbre. Naturally produced stimuli from the piano, harpsichord, and brass instruments are all characterized as having a quick attack and rapid decay, whereas woodwinds and strings have a relatively slow attack and gradual decay (Fletcher, 1991). The piano, flute, violin, and trumpet were chosen because these instruments are readily discriminable from one another. The harpsichord and piano served the role as possibly confusable instruments because of their similar timbres that may be based upon attack, partials, or overall waveform properties. If the harpsichord and piano are indeed easily confusable, one should expect to find faster reaction times for comparisons across these instruments relative to comparisons across discriminable instruments.

There were two 871.5 ms samples of each instrument recorded for each of four triads: C-major (C-E-G), C-minor (C-E♭-G), E♭-major (E♭-G-B♭), and E♭-minor (E♭-G♭-B♭), where possible, in the same octave above middle C[1]. This ordered chord progression represents the degree of relatedness for any two given chords, with each listed chord differing from the immediately preceding chord by a factor of one note: i.e., C-major differs from C-minor by one note, from E♭-major by two notes, and E♭-minor by all three notes. The chords were recorded on a high-bias chrome cassette, then converted to a 12-bit digital representation at a 10 kHz sample rate with 4 kHz low-pass filtering. A 386-DOS computer was used on-line to randomize trial order, present the stimuli, time events, and record responses. The stimuli also were 4 kHz low-pass filtered at presentation, and were delivered binaurally over TDH-49 headphones in a commercial acoustic chamber.

Procedure. Experiment 1 used an AX chord discrimination task that was blocked for instrument condition. Subjects were instructed to judge whether or not the two chords presented on a given trial were equivalent (consisted of the same notes), with emphasis placed on the need to disregard the instrument playing the notes. Following the procedures typically used in normalization studies, instrument variability was manipulated within subjects by presenting stimuli in two blocks representing unique instrument conditions: Single Instrument, and Mixed Instrument. The 120-trial Single Instrument condition consisted of stimuli from only a single instrument within each trial, but stimuli from different instruments across trials. The Mixed Instrument condition consisted of chords played by different instruments both within and across the 480 trials. The order of conditions was counterbalanced across subjects. Subjects were informed of the exact nature of each condition prior to the block of trials. Within each condition there was an equal (0.50) probability of same and different chord presentations.

A trial consisted of presentation of the A stimulus, a 1500 ms ISI, the X stimulus, and a 3000 ms response interval. Subjects were instructed to respond as quickly and accurately as possible; responses were indicated by pressing one of two keys (corresponding to same vs. different chord) on a response pad. All responses were recorded by the computer that measured RT using a 1 ms time-base and prohibited any change in response.

Results and Discussion

Based upon speech normalization findings, subjects should be faster (and possibly more accurate) in the Single Instrument condition compared to the Mixed Instrument condition. Table 1 shows mean accuracy and RT results. The mean RT values across subjects were obtained from individual median RT scores for correct responses only. The full set of RT and accuracy data were subjected to separate 2 x 2 ANOVAs, with instrument condition (Single vs. Mixed) and chord (same vs. different) serving as variables. The analysis of different chord comparisons was further broken down by the number of notes between chords; discussion of this finer analysis of results will follow the separate discussion of general RT and accuracy results.

Insert Table 1 Here

The results in Table 1 show that subjects performed quite accurately in the Single Instrument condition and, in comparison, very poorly in the Mixed Instrument condition [$F(1,19) = 164.78$, $p < 0.01$]. This effect of instrument condition occurred for both same and different chord trials [$F(1,19) = 134.03$, $p < 0.01$, and $F(1,19) = 65.20$, $p < 0.01$, respectively]. The reduced accuracy [on both same and different chord trials] for judgments across instruments indicates that subjects were seldom able to ignore irrelevant stimulus differences associated with instrument timbre. Although performance on single instrument trials was significantly faster than mixed instrument trials [$F(1,19) = 17.85$, $p < 0.01$], as hypothesized, the accuracy data appear to indicate a general failure to normalize, and therefore do not allow the use of RT to evaluate differences in processing (i.e., normalization). However, two subjects performed at relatively high levels of accuracy under both Single and Mixed Instrument conditions, and their data (which are discussed later) allow some evaluation of normalization for instrument timbre.

There was no main effect of same/different chord on accuracy [$F(1,19) = 0.48$, $p < 0.10$], but there was a significant chord by instrument condition interaction [$F(1,19) = 17.66$, $p < 0.01$]. In the Single Instrument condition there was a significant tendency to respond "same" [$F(1,19) = 14.38$, $p < 0.01$], whereas in the Mixed Instrument condition there was a tendency to respond "different" [$F(1,19) = 2.43$, $p < 0.14$]. Thus, it appears that subjects may have responded as much on the basis of overall timbre as on the basis of equivalence of chords.

Chord Analysis. Table 2 also summarizes the effects of note differences for different chord trials across Single and Mixed Instrument conditions. Discrimination of one-, two-, and three- note differences between chords represents increasing S, and thus increasing S/N. Accuracy should be expected to increase, and RT to decrease, with greater S/N. Thus, it is not surprising that performance was better for discrimination of three note differences than one- or two-note differences, but the absence of any differences in RT was not expected. The lack of note difference effects on RT may have been due to the high error rates in the Mixed Instrument condition. There was a marginal main effect of same/different chord on RT [$F(1,19) = 3.95$, $p < 0.10$], which is attributable to the significantly slower response times for different chord trials [$F(1,19) = 5.98$, $p < 0.05$] in the Single Instrument condition (where accuracy was higher).

Instrument Analysis. Table 2 shows d' as a measure of accuracy for each instrument combination on the AX trials. d' was computed *for each listener* based upon probabilities that subjects responded "same" to equivalent (hit) and nonequivalent (false alarm) chords. d' then was averaged across subjects (Pastore & Scheirer, 1974). Table 2 also shows RT for the instrument comparisons.

Insert Table 2 Here

Based upon the summarized research on instrument timbre, it was expected that the similarity in attack, and possibly upper partials, of the stimulus waveforms would have differential effects on response time. For example, a piano-harpsichord comparison (where both instruments have quick attacks) should have been faster and more accurate than a piano-woodwind comparison (where woodwinds have relatively slow attacks).

For both same and different chord trials, d' was always highest, and RT almost always fastest for single instrument comparisons. However, none of the expected effects based on instrument waveshape were found. One possible explanation for the lack of instrument similarity effects may be that although the instruments were selected based upon expected differences in attack, attack may not have been extremely critical to distinguishing timbre in the tokens used. It is also possible that subjects were not sufficiently experienced with music stimuli to effectively utilize specific timbral components such as attack; the lack of normalization for most subjects is consistent with this possibility. Also, in mixed instrument comparisons, d' was higher and RT faster (with a few exceptions) when the second stimulus was played by the piano. It may be that subjects, in general, have more exposure to piano timbre or chords played on the piano (compared to other musical instruments), and thus are more efficient perceivers of this particular timbre. Other work from our laboratory is consistent with this conjecture (Hall & Pastore, 1993).

Testing Musically Proficient Subjects

Experiment 1 sought to demonstrate normalization in terms of increased response latencies, but not decreased accuracy, in conditions where timbre differed within trials. Subjects performed quickly and accurately on single instrument chord comparisons, and significantly slower, but also with considerably less accuracy, on mixed instrument chord comparisons. These overall results have not produced support for a meaningful normalization process for music timbre. Although the subjects were asked to self-select for participation based on musical experience, the accuracy results suggest that the majority of the participants were not proficient musicians.

The results from Experiment 1 are consistent with suggestions that chord and timbre may be integral [in the Garner (1974) sense]. If two dimensions are truly integral, then subjects should not be able to normalize for one dimension when perceiving the second dimension. For example, Krumhansl and Iverson (1992) found that pitch (of isolated tones) and timbre do interact to some degree (or are not perceived independently); subjects could not attend to the pitch of a tone without being influenced by its timbre. Musical experience may be a factor in the degree of integrality between pitch and timbre. Wolport (1990) obtained results that suggested timbre is more salient than pitch for nonmusicians than musicians; alternatively, it seems possible that nonmusicians may have the ability to separate chord from timbre, but have difficulty in understanding what is required of them. Consistent with both types of explanations, Beal (1985) reported that nonmusicians found it more difficult

than musicians to judge two chords as the same when they were played on different instruments. There also are a number of other reports (e.g., Pitt, under review) that nonmusicians have difficulty separating pitch and timbre. Thus, music experience may be important in the degree to which pitch and timbre interact.

Could the results of Experiment 1 have been due to the lack of reasonable musical experience for the subjects, or do the results instead reflect limits on music perception processes that occur even for experienced listeners? Separate analyses of data for musicians and nonmusicians may be required to investigate this possibility that music experience differentially affected timbre and pitch perception.

Musical histories were known for 4 of the 20 subjects in Experiment 1. Two (Subjects 7 and 13), were proficient musicians and performed very accurately in all conditions. Their data thus allow some evaluation of normalization. Subject 21 had moderate musical experience, although to a lesser extent than Subjects 7 and 13, whereas Subject 2 had only about 1 year of music experience.

Table 3 provides accuracy and RT data for the 4 subjects. Subject 7 performed at high levels of accuracy throughout the experiment, with 100 and 96 percent correct on same chord trials for Single and Mixed Instrument conditions, respectively. The RT data for this subject are somewhat consistent with normalization predictions, with increased variability in the different instrument conditions resulting in longer response times for same chord trials [$t(9) = .1.78$, $p > .10$]. Thus, it appears that Subject 7 may have normalized for stimulus variability, maintaining accuracy at a cost to speed. Subject 13 also performed at 100 percent correct for the same instrument conditions and at a good, but somewhat poorer level of 92 percent correct on different instrument conditions for same chord trials. Although RT was in the expected direction, the difference did not approach significance [$t(9) = .84$, $p > .50$]. Subjects 2 and 21 performed at 97 and 100 percent correct, respectively, for Single instrument trials. However, performance for these subjects was poorer for the different instrument trials, at 55 and 68 percent correct. The RT data for these two subjects do follow the expected normalization patterns, where RT for different instrument - same chord trials were significantly longer than same instrument - same chord trials [$t(9) = 3.52$, $p < .01$, and $t(9) = 5.69$, $p < .002$ for Subjects 2 and 21, respectively]. Based on the results from our musically experienced subjects (7 and 13), it is possible that testing clearly experienced musicians may resolve the accuracy problem and provide stronger support for an active model of normalization. This was the primary goal of Experiment 2, which examines normalization as a function of musical proficiency.

Insert Table 3 Here

Experiment 2

The logic behind the second experiment begins with the assumption that normalization is an active process. When a comparison is to be made between two sequentially presented stimuli, the first stimulus then should set up expectations about the parameters or nature of the processes to utilize in analyzing the second stimulus. Furthermore, the first stimulus need not be auditory to generate expectations about a subsequent stimulus. Presentation of a visual cue for instrument should provide sufficient information for the system to anticipate well-learned stimulus characteristics. One would expect valid cues (which correctly cue expectations about the instrument playing the auditory stimulus) to produce faster response times than invalid visual cues (which set the processes for an incorrect instrument). If normalization instead is a passive process, then expectations become irrelevant for stimulus processing. Thus, the nature of visual cues should not influence response times.

Following this logic, Experiment 2 used a Posner (1980) cross-modality cuing paradigm to determine whether the perceptual system can actively anticipate, or can only passively process, irrelevant stimulus (timbre) variability. Subjects with known musical abilities were tested for reaction time effects with relatively high levels of performance accuracy. Since the visual cues should not directly alter the S/N ratio, there should be no effect of valid or invalid visual instrument cues on subsequent chord judgments if normalization only reflects a decrease in S/N. However, if normalization is an active process, then longer reaction times would be predicted for invalid cues, reflecting the perceptual system preparing for irrelevant (timbre) variability and/or determining that the expectations were inappropriate.

Method

Subjects. Six subjects participated in Experiment 2. With the exception of Subject 6, who had only studied an instrument (piano) for approximately 1 year, most subjects had over 5 years of music experience. The 4 subjects with known musical histories from Experiment 1 also participated in Experiment 2. All subjects reported normal hearing.

Stimuli and Procedure. The experimental design was a speeded, single-stimulus, major/minor chord labeling task, and employed a cuing procedure adapted from the work on visual attention by Posner and colleagues (Posner, 1980; Posner, Snyder, & Davidson, 1980). Subjects were presented with a 500 ms visual cue followed by an ISI that randomly varied between 1 and 2.5 s in 500 ms increments. Based upon the visual work, the variable ISI was intended to eliminate general, temporally-related, anticipatory effects due solely to the presence of the visual cue. The ISI was followed by a C-major or C-minor chord from Experiment 1 played by one of four instruments: piano, harpsichord, brass, or strings. The woodwind stimuli were not used in Experiment 2 because subjects in Experiment 1 often reported that the instrument sample did not sound characteristic of its natural counterpart.

The subjects' task was to identify each triad as major or minor by differential button presses on a keypad. Cues, which were always orthogonal to (and conveyed no information about) whether a chord was major or minor, were neutral, valid, or invalid with respect to the instrument playing the target. The cue "+ + + + +" was neutral with respect to instrument and was presented with a probability of 0.17. Instrument cues were four-character representations of a given instrument ["STRG" (strings), "PINO" (piano), "BRAS" (brass) or "HARP" (harpsichord)], and occurred on the remaining trials ($p = 0.83$).

Instrument cues, when presented, were valid with a probability of 0.80, and thus were valid with an overall probability of 0.66 ($p = 0.80 * 0.83$) across all trials; instrument cues were invalid with an overall probability of 0.17 ($p = 0.20 * 0.83$). Three blocks of 200 trials were presented.

Results and Discussion

Figure 1 shows mean reaction times obtained from individual median scores for correct responses given valid, neutral, and invalid cue trials. In the figure subjects are ordered in terms of increasing RT for neutral cue conditions. Standard error bars also are provided. A one-way ANOVA of RT revealed a significant main effect of cue validity [$F(2,5) = 6.58$, $p < 0.05$]. This effect was primarily a result of consistently longer reaction times to invalid trials relative to valid trials (Tukey-test $p < 0.50$). Response latencies for neutral trials also were consistently longer relative to those for valid trials, but this difference did not reach statistical significance (Tukey > 0.05).

Insert Figure 1 Here

There were minimal differences in accuracy effects, with the two most highly practiced musicians (Subjects 1 and 3 in the current experiment, and subjects 7 and 13 respectively, in Experiment 1) performing at essentially 100 percent correct, and with Subjects 2, 4, and 5 performing at better than 90 percent correct regardless of cue and target instrument. These ceiling effects were not viewed as a problem because we sought to demonstrate an active normalization process through RT effects in the context of consistently high accuracy levels. Subject 6 (who was a relatively inexperienced musician and who had the least amount of training of the 6 subjects) performed at approximately 79 percent correct and also was the only subject whose responses on valid cue trials were not significantly faster than invalid trials.

The relatively accurate responding of subjects across all conditions is consistent with Krumhansl and Iverson's (1992) conclusion that "the interaction between timbre and pitch of single tones does not imply that it is impossible to abstract and compare pitches of tones with different timbres or timbres of tones with different pitches... The increased reaction times in tasks requiring this information to be abstracted indicates that this process requires additional time (p.749)."

It was anticipated (as in Experiment 1) that the relationship between the expected properties of cued instrument waveform and the perceived properties (based on waveshape) of target instruments would influence RT and accuracy. Table 4 reports mean reaction times for each possible combination of cue and target instrument. Although no consistent effects of waveshape were obtained, there were some notable trends in the RT data. It appears that RT was fastest for the piano in both valid and neutral cue trials. For 8 of the 12 invalid cue comparisons, reaction times for trials where both instruments had a rapid attack were faster than trials with slower attack instruments. For instance, the piano-harpsichord comparison was faster than the harpsichord-string comparison. The results also show that the harpsichord-piano comparisons were, by far, the fastest of all invalid cue trials. These results might be due to the similar attack functions of the piano and harpsichord, but might also be due to a combination of high familiarity with the piano as a chord instrument and to similar overall waveshape properties between the harpsichord and piano. The fast reaction times for harpsichord-piano comparisons also provide an important basis for later experiments which investigate the role of perceptual similarity in normalization; based upon an implicit assumption from demonstrations of speaker normalization (Logan, 1990), similar tokens should be processed faster than dissimilar tokens.

Insert Table 4 Here

The results of Experiment 2 clearly provide evidence that normalization is an active process. Subjects were significantly slower when the visual cue was invalid than when the cue was either valid or neutral, indicating a cost for inappropriate perceptual expectations. This cost is an increase in processing time due to inappropriate expectations, resulting in a need to identify the inaccurate perceptual setting, then to readjust for proper source variability.

The use of digitally sampled, rather than natural instrument tokens may have limited the observance of consistent waveshape effects. Subjects in both experiments often reported that the harpsichord samples did not sound characteristically like the natural instrument; such unnatural timbre had already been identified as a problem for the woodwind samples in Experiment 1, and was the basis for not using those stimuli in Experiment 2. Therefore, it is possible that some stimulus properties contributing to timbre in our synthetic tokens may have prevented consistent normalization to expected stimulus properties that otherwise could have been in evidence. Thus, such additional properties may have limited the already significant main effects of cuing in Experiment 2. This possibility will be investigated in Experiment 3 by using natural instrument tokens. In addition, waveform manipulations of natural tokens may help identify the salient components of timbre, which should, in turn, be important for normalization.

Experiment 3

If normalization is to be considered an active process, a pertinent question becomes what physical stimulus properties are actually being factored-out or normalized? The goal of Experiment 3 was to identify global stimulus properties that contribute to the characteristics of particular instrument timbres, and thus form the basis for perceived instrument variability that may be subject to normalization. A number of studies (see the introduction) have indicated that attack functions, and possibly spectral composition, are largely responsible for instrument timbre. Experiment 3 provides an evaluation of the importance of attack and spectral composition (in terms of upper partials) in defining the similarity (or conversely, the distinctiveness) of instrument timbres.

Natural instrument tokens were physically altered to evaluate possibly critical components in instrument timbre. Two

types of physical alterations were applied to the natural stimuli: (1) removal of the attack portion of each stimulus, and (2) removal of all higher-order partials. Experiment 3 thus utilized 3 sets of "natural" stimuli: 1) full (intact) versions of the natural tokens, 2) tokens with the attack portions removed ("cut-attack"), and 3) filtered tokens. The manipulated stimuli represent extremes, in that attack transitions and most partials were completely eliminated. Thus, if the attack and/or upper partials play significant roles in instrument timbre, eliminating the given property should result in stimuli which are highly similar to each other and dissimilar to the original, unaltered stimuli.

A similarity scaling procedure was used to evaluate the importance of the attack and upper partials as components of instrument timbre. Subjects were randomly assigned to one of three conditions. In Condition 1 the synthetic stimuli from Experiment 1 were compared to their intact natural counterparts. This comparison allowed us to assess the generalizability of the Experiment 2 results obtained using synthetic stimuli to natural instrument tokens. In Condition 2, natural instrument tokens with the attack portions removed were compared to intact natural stimuli. If the attack component of a chord is a significant component of timbre, instrument tokens with missing attack transitions should be perceived as highly similar to each other and not very similar to intact versions of the same instrument. In Condition 3, low-pass filtered natural tokens were compared to natural intact stimuli. Using the logic from Condition 2, if the higher-order partials are significant components of timbre, filtered stimuli should be perceived as highly similar to each other and not very similar to intact versions of the same instrument.

Method

Subjects. Forty-two students enrolled in psychology courses at Binghamton University served as subjects (14 subjects for each of the 3 conditions) in partial fulfillment of course requirements. All subjects reported normal hearing and had at least 2 years of musical experience.

Stimuli. The stimuli consisted of chords produced by five natural instruments. The instruments were the same as those used in Experiment 1. A practiced musician with a given instrument produced the isolated notes C, E, $E^{\flat}$, and G, each approximately 870 ms in duration. The samples were recorded on a half-track tape recorder (Tandberg TD 20A operating at 15 ips) using a B&K Model 4135 microphone. The recorded stimuli were digitized (12-bit with 10 kHz sample rate and 4 kHz low-pass antialiasing filter). The stimuli were then digitally mixed to produce C-major chords using a 386-DOS computer. Thus, all chords were created as though three musicians simultaneously played the component notes, even if the full chord could have been played on a single instrument (e.g., piano or harpsichord).

The cut-attack stimuli consisted of intact stimuli which were digitally edited to remove the attack functions. The most intense portion of each token was identified to determine the length of the attack function. The attack (the stimulus portion prior to and including peak amplitude) then was excised from the waveform at a waveform zero-crossing. In order to prevent any sudden onset transients, each cut-attack stimulus then was amplitude-weighted by imposing a constant, brief (30 ms) linear onset.

The filtered stimuli consisted of intact chords that were low-pass filtered at 500 Hz, with a 72 dB/octave skirt. For these stimuli, most of the higher-order partials were removed, leaving the fundamental plus, at most, one partial from the C note and, possibly, one partial from the E note. Individual tokens were attenuated to equate overall peak amplitude across stimuli. Time varying spectrograms for each of the stimuli were examined to confirm the effectiveness of the temporal and spectral manipulations performed on the stimuli. Spectrograms showed that the onset portion of each waveform was absent for all the stimuli in the cut-attack manipulation. Similar analyses for the filtered stimuli showed that most of the higher harmonics were removed. Obviously, both manipulations eliminated the relative onsets of the upper partials, which is a potential interactive cue for timbre.

Procedure. Subjects were instructed to judge the similarity of the two stimuli presented on each trial. A 7-point scale was used, with "1" indicating minimal similarity and "7" indicating maximal similarity. Each condition was composed of three parts. In Part 1, all stimuli in the experiment were presented sequentially to give the subjects an idea of the range of the different instrument tokens. In Part 2, subjects were given an example of a very dissimilar and a very similar pair of stimuli. Different stimulus pairs were used as examples, depending on the condition. These examples were not expected to produce demand characteristics since the subjects were told that the pairs should not be considered as the maximally similar or dissimilar comparisons on which to base their later judgements, but rather, just as examples of similar and dissimilar items. No data were collected in these initial familiarization sections of the experiment. In Part 3, similarity ratings were collected on each of 450 trials. A trial consisted of the first stimulus, a 1500 ms ISI, and presentation of the second stimulus. Subjects were asked to respond within a 3000 ms response interval by pressing keys corresponding to 1-7 on a response pad; these responses were recorded by the computer

Results and Discussion

For every condition, the average similarity rating was calculated across subjects for each of the 90 stimulus pairs. The mean similarity ratings were submitted to a multidimensional scaling program (Systat) using a Euclidean metric (Minkowski metric with $r = 2$), which is appropriate for integral dimensions (Garner 1974). [Mean ratings were also analyzed using a city-block metric ($r = 1$) which is appropriate for separable dimensions. Since both metrics yielded highly similar solutions, only the Euclidean solution is presented and discussed.]

For Condition 1, where full versions of the natural and synthetic stimuli were compared, a 2-dimensional solution was obtained (stress value of 0.054). Panel A of Figure 2 shows the MDS space based on dimensions 1 and 2. Table 5 provides a decoding of the instrument labels used in the figure. With the possible exception of the harpsichord and woodwinds (flute), where similarity still is quite high, overall similarity between the natural instruments and their synthetic counterparts was very high.

Insert Figure 2 and Table 5 Here

Dimension 1 was related to the resonant properties of the instruments; in other words, the wind instruments (flute/woodwind and trumpet/brass, whose sounds emanate from a tube) were grouped together, while the string instruments (violin, piano, harpsichord) also occupied a similar space. Dimension 2 appeared to be related to degree of spectral fluctuation. For example, flutes tend to have upper harmonics that rise in amplitude (at onset) and decay (at offset) in close alignment, whereas strings and brass instruments (which grouped separately along Dimension 2) tend to have varied amplitude patterns for individual partials at onset and offset (see Grey, 1977).

With the exception of the woodwind (flute) and harpsichord (both noted as unnatural stimuli in Experiments 1 and 2), the synthetic instrument samples appear to have been adequate examples of their natural instrument counterparts. Table 6 verifies this assertion by providing mean similarity ratings for comparisons of synthetic and natural instrument counterparts. Although the synthetic and corresponding natural tokens were generally very similar, the finding that they were not maximally similar to each other suggests that there are at least some subtle, but perceivable differences between the two types of stimuli. This finding indicates that the results from Experiments 1 and 2 are reasonably valid, but also support our rationale for using natural instruments in the remainder of this and in the last experiment.

Insert Table 6 Here

Condition 2 (natural intact vs. natural cut-attack) resulted in a 2-dimensional solution (stress value of 0.068) which is shown in panel B of Figure 2. Each cut-attack stimulus was highly similar to the intact version from which it was derived. In fact, the trumpet (Tn and Tc) stimuli were maximally similar to each other. Based upon the position of the groupings in the perceptual space, Dimension 1 appeared to be related to resonant properties of the instruments, while dimension 2 was related to degree of spectral fluctuation (both summarized above).

It did not appear that the cut-attack manipulation of the stimuli had any significant effect on similarity compared to their intact counterparts. Panel A of Figure 2 (which illustrates the MDS space for the intact stimuli) and panel B of Figure 2 (which depicts the MDS space for intact and cut-attack stimuli) show approximately equivalent similarity spaces for the intact and cut-attack tokens, with the elimination of the attack not meaningfully affecting the similarity of the altered stimulus to its original counterpart. This finding is contrary to the conclusion by Grey that onset functions are important in instrument identification. Although it is possible that some aspect of the synthetic nature of Grey's stimuli may have contributed to the greater importance of attack noted in his study, it is unlikely given that his stimuli were based directly on natural tokens using an analysis-by-synthesis approach. A more likely reason for this difference in results could be related to the lengths of the stimuli, as suggested by Handel (1989) Grey's stimuli, which were synthetic single tones, ranged from 250-500 ms, whereas in the current study the natural chord stimuli were approximately 870 ms. The onset transitions in the current study thus could have constituted a much less significant portion of the stimuli than those used in the Grey study. If this conjecture is valid, then removal of the attack functions should lead to significant decrements in instrument identification for shorter stimuli. We will address this argument more completely in the general discussion.

Condition 3 (natural intact vs. natural filtered) also resulted in a 2-dimensional solution (stress value of 0.048). Panel C of Figure 2 provides the MDS space for the filtered and intact stimuli based on dimensions 1 and 2. Dimension 1 was related to the presence/absence of higher overtones of the tokens; the filtered stimuli were all highly similar to each other and minimally similar to their intact versions. Dimension 2 was related to the resonant properties of the instruments (as discussed above). Based upon spectrograms of the intact tokens of the flute stimuli, which verified that the flute is characterized by very weak higher harmonics, the filtering procedure was not expected to (and did not) have a dramatic effect on timbre.

Experiment 4

In order to better establish the relationship(s) between listener expectations and physical characteristics of stimuli, Experiment 4 used the intact and the two sets of physically-altered stimuli in a major/minor chord discrimination task. If some type of normalization process (that is based upon adjusting for differences between two stimuli) is indeed invoked, instrument comparisons that resulted in low similarity ratings should result in increased reaction times. Conversely, instrument comparisons that were judged to be very similar should result in a minimal increase in S/N and thus, a minimal need for normalization, as demonstrated by faster reaction times. The results of Experiment 3 will directly evaluate the roles of the described global timbral properties (derived in Experiment 3) in normalization.

Method

Subjects. Twelve subjects from Binghamton University, each with at least 5 years of music experience, served as subjects for this experiment. All subjects reported normal hearing and were paid $5 per hour for their participation.

Stimuli and Procedure. C-major and C-minor chords were used as stimuli. Experiment 4 used a chord discrimination task similar to the one used in Experiment 1. Full, intact stimuli were paired with full, cut-attack, and filtered stimuli from the same instrument or from different instruments. In order to limit both the number of total trials and stimulus uncertainty, the first chord presented was always a C-major chord. This standard stimulus was followed, after a 1500 ms ISI, by either a C-major or C-minor chord. The experiment consisted of 720 trials with brief rest periods provided between each block of 120 trials. Subjects were instructed to respond as quickly and accurately as possible. A maximum of 3000 ms was allowed to press one of two keys, corresponding to judgments of "same" and "different" chord, on a response pad. All responses were recorded

on-line by a computer that measured RT with 1-ms accuracy.

Results and Discussion

Subjects performed at high levels of accuracy, averaging 94 percent correct. Reaction times were obtained from individual median scores for correct responses only. A linear regression analysis was performed on the scaling data obtained from Experiment 3 and reaction times for the corresponding stimulus pairs from Experiment 4 to demonstrate possible systematic changes in RT as a function of stimulus similarity. The linear regression line resulted in an r^2 of 0.532, and indicated that as similarity increases there is a corresponding, and fairly consistent, decrease in RT. It appears that this significant correlation would have been higher were it not for two outlying scores. Both cases involved comparisons of a full stimulus followed by a cut-attack stimulus. Reaction time for these comparisons were faster than what was predicted. Since these two comparisons were different instrument conditions, slower reaction times had been predicted. It is suspected that the cut-attack manipulation, when presented as the second stimulus, may have speeded RT to some extent because peak amplitude would be reached at a much earlier point for these stimuli. Thus, critical information necessary to identify a certain timbre may have been obtained earlier by the listener.

To investigate the possibility that subjects were able to respond more quickly to the cut-attack stimulus, another brief, additional chord judgment experiment (similar to Experiment 4) was conducted, using only the full and cut-attack stimuli and manipulating the order of presentation. On any given trial the cut-attack stimulus could be presented as the first or second stimulus. As suspected, a one-way ANOVA revealed that there was a significant predicted effect of order of presentation [$F(3,4) = 6.28$, $p < 0.01$]. Based on this information, a second linear regression was computed using the original data, but omitting the two conditions where a cut-attack instrument token was presented as the first stimulus. Figure 3 shows the results of the second regression, where the new value of r^2 (0.689) is indeed higher than that obtained in the first regression.

Insert Figure 3 Here

The results of Experiment 4 demonstrate that timbral similarity of two items is an important predictor of processing time for normalization. Furthermore, timbre is primarily dependent on the information that is present in the upper harmonics of instrument tokens (Experiment 3). Increases in processing time are reflective of the degree to which timbral differences between stimuli are factored-out. Highly similar tokens are processed faster since there is less variability to adjust. Correspondingly, additional time is required to normalize greater stimulus variability.

General Discussion

The current investigation has shown that normalization processes for task-irrelevant source variability are not unique to speech. Thus, the present nonspeech finding of timbre normalization in chord identification suggests that normalization may reflect a general auditory perceptual mechanism. We note from experience (e.g., Hall & Pastore, 1992) the difficulty in providing a strong empirical evaluation of claims concerning whether of not speech is mediated by a specialized, biological mechanism for processing speech (Liberman & Mattingly, 1985, 1989; Whalen & Liberman, 1987). We therefore simply acknowledge attempts of previous normalization studies (Pisoni, Carrell, & Gans, 1983) to address that issue, and focus our discussion on implications of normalization for the nature of perceptual processing.

Implications for Perceptual Processing

The perceptual system appears to have the ability to factor out, or at least partition, information associated with irrelevant features prior to complete identification of the relevant features. By providing a demonstration of such partitioning of information, the current series of experiments provides some important insights into general aspects of perceptual processing, particularly in terms of attention to specific features.

Experiment 1 showed that variation in timbre results in significant performance decrements for both accuracy and RT. These results may reflect a failure on the part of most subjects to effectively normalize, with only two highly practiced musicians demonstrating a cost to RT without a decrement in accuracy when instrument was varied (see "Effects of Music Experience", below). Thus, normalization may reflect the use of acquired knowledge in an efficient, possibly automatic fashion. Experiment 2 demonstrated that there is an active, anticipatory component to normalization rather than being a phenomenon solely based upon a passive response to increased stimulus variability (or decreased S/N ratio). In this active process, it appears that the perceptual system does not simply evaluate each attribute by combining the values of the limited set of relevant features, but rather seems to actively engage first in setting some processing parameters (based on expected stimulus properties), and then evaluating the adequacy of the setting. If the settings are incorrect, the system can modify the setting, but with a loss of time. This loss of time is reflected in a greater cost for an invalid cue than an advantage for a valid (relative to a neutral) cue: this asymmetry of cost to benefit has been reported often by Posner (1980) for visual stimulus processing. Therefore, normalization has been demonstrated to be an active, adaptive type of stimulus processing. By the same token, it may be reasonable to conjecture that limits of time or processing capacity would result in decreases in accuracy resulting from the perceptual system resetting processing parameters or utilizing a more global processing strategy, either of which should result in an increase in N, and thus a decrease in S/N ratio.

The perceptual system may always perform a survey of the stimuli, and determine the appropriateness of the existing setting/algorithm (from a previous trial or from a cue). If the system has determined the necessity to change the normalization algorithm, there may be a cost in time and processing resources. Prior to activation of the appropriate algorithm, there must be a disengagement of any inappropriate algorithm. The notion and cost of a disengagement process also has been described in the visual attention literature by Posner (1980; also see Experiment 2)

Conceptualizing normalization as a central form of adaptive processing provides some possible accounts for the significant RT cost for invalid cues in Experiment 2. First, in the initial appraisal of the stimuli, the system should be able to perceive dissimilar stimuli more quickly, activating a faster change in setting, and resulting in a minimization of response time. However, our empirical results refute this possibility. Second, for highly similar stimuli, the system may retain the existing setting (i.e., there is no normalization), with some added noise due to even small differences in the stimuli. In this case, slower response times and decreased accuracy would be expected, however, the systematic changes in RT as a function of similarity that were found would not be predicted. Third, the time to disengage an algorithm should not differ as a function of similarity. This possibility is also refuted by the results of Experiment 4, where reaction time is inversely related to similarity. Thus, the most plausible explanation of the normalization process is that time/effort for adjustment to an appropriate algorithm or setting *is a function of similarity. Perceptually similar tokens* are processed faster since there is less variability to partition or factor out. Conversely, greater time and effort is required to normalize larger stimulus variability.

Effects of Music Experience

The degree to which algorithms are effectively utilized during normalization seems to be a function of the amount of experience with a particular class of stimuli. For example, performance was faster and more accurate *in conditions that included* a piano stimulus, and many of our subjects were pianists. Possessing greater formal knowledge of the theoretical structure/underpinnings on which the stimuli are based may be an important factor in stimulus processing in timbre normalization.

Highly practiced subjects should have available efficient algorithms to normalize for the effects of instrument variability. A highly related alternative to this conceptualization is that musicians may use different types of knowledge (that nonmusicians may not possess) to invoke imagery or schema-based representations in performing certain tasks. Most subjects in Experiment 1 tended to have limited musical experience. These subjects thus may have had relatively inefficient, and possibly inaccurate, normalization algorithms that probably functioned more on immediate experience with stimuli than on knowledge of chords and instruments. Following this conceptualization, speed and accuracy could have been superior in the Single Instrument condition because the subjects applied the same (possibly incomplete and inaccurate) normalization algorithm to the A and X stimuli, resulting in equivalent normalization errors for A and X stimuli. Those same errors would not have been equivalent for A and X stimuli in the Mixed Instrument condition, where different algorithms must be applied to stimuli. The result would be an increase in the perception of differences in the pitch attribute, and the observed increase in errors in the Mixed Instrument condition.

Comparisons with Imagery

The normalization process might well involve some form of auditory imagery or schema. For example, Subject 1 *reported anticipating a rapid stimulus* onset when cued for the piano. Subject 5 (Experiment 2) also reported attempting to anticipate the complete chord played by the cued instrument. These anticipations are consistent with the use of some form of imagery, and may reflect the typical nature of expectancies for at least some subjects in the music normalization process.

An excellent study by Crowder (1989) on auditory imagery for timbre used the same basic strategy as our first two experiments, but with somewhat different goals and results. Experiment 1 of the Crowder study attempted to demonstrate the effects of instrument variability on pitch judgments for single tones, establishing a basis for demonstrating positive and negative effects of imagery in the second experiment. Crowder obtained a main effect of instrument (with significantly faster RTs on same instrument trials), thus, in effect, demonstrating normalization. However, these normalization results were limited by a significant instrument by pitch interaction, such that the main effect of instrument was observed only on same-pitch trials. Thus, as with our normalization results (e.g., Experiment 1), instrument variability resulted in increased response latency. Crowder also had some problems with subject accuracy, as we did in our Experiment 1. Data from 3 subjects in the Crowder study (who performed significantly below chance) were discarded in order to obtain the predicted *effects of timbre variability.*

In his Experiment 2, Crowder used a cuing technique that also involved a self-paced AX task (the fixed trial structure in our Experiment 2 used a single interval identification task with a visual cue). The subjects were instructed to imagine the presented sine tone being played by a certain instrument. After the subjects had indicated the formation of an image, an instrument tone was presented, whereupon a "same/different instrument" judgment was made. Crowder obtained a similar interaction of pitch and imagined timbre to that found in his first experiment, with the expected effect of instrument variability observed only for same-pitch trials. Thus, it appears that subjects can actively image timbre, with processing costs (slower RT) for (imagined) properties that are not consistent with subsequently presented *stimuli. These results are compatible* with the normalization findings in our Experiment 2, where subjects appeared to actively engage in setting parameters based on expected stimulus properties. Thus, although our subjects in Experiment 2 were not specifically instructed to use auditory imagery, our results are consistent with evidence for the use of timbral imagery. In contrast to our Experiment 2, the subjects in Experiment 1 need not have actively generated from memory an internal representation of an expected auditory stimulus, but rather may have compared a trace (or echoic image) of the first stimulus with the second stimulus.

Timbral Contributions of Spectral Characteristics and Attack

Experiment 3 demonstrated that upper harmonics, but not attack functions, play a significant role in the timbral characteristics of instruments, and thus should be most subject to normalization processes in the discrimination of chords. Similarly, Pitt and Crowder (1992) demonstrated an inability to image rise time (loudness), a salient component of attack functions, and concluded that timbral imagery is based primarily on spectral properties.

The role of dynamic onset properties in timbre representation may be a function of the length of the stimulus (e.g., see Handel, 1989). For example, in order for the attack functions to influence performance in any AX task (as in our Experiments 1, 3, and 4), subjects must compare the X stimulus attack with some form of representation of the A stimulus

attack. This representation could be a trace (or type of echoic image), or, alternatively, an encoded version of the attack. These possible representations are respectively equivalent to what have been called "trace coding" and "context coding" processes (e.g., Macmillan, Braida, and Goldberg, 1987).

A trace memory will decay rapidly over time. The vast majority of estimates suggest that trace decay should be complete within 1-2 s (e.g., Darwin, Turvey, and Crowder, 1972; Treisman and Rostron, 1972). As ISI approaches or exceeds this 1-2 s limit, context coding becomes the more viable strategy.

Let us temporarily assume that our approximately 1 second stimuli include a 100 ms attack portion. With an ISI of 1.5 s, the functional delay for comparing the attack function of A and X stimuli therefore becomes [(1s-100 ms) + 1.5 s], or 2.4 s. Thus, even in the absence of masking from the final portion of the A stimulus (which also may occur), an adequate trace of the attack function of the A stimulus will no longer be available for comparison with the X stimulus. The only attack information for the A stimulus thus must be some form of context coding.

There also are several reasons to expect that context coding of attack information be poor in longer stimuli. For example, if context coding is capacity limited, then encoding should be best for simpler, more salient, stimulus properties. In longer stimuli, like those used in the current experiments, long-lasting, static information is (generally speaking) consistently available after a relatively brief and more complex attack. Thus, following the logic based on limited capacity, static properties (rather than attack properties) should be encoded.[3] Furthermore, if context coding requires processing time to access memories for particular stimuli, then encoding of stimuli must be weak in instances where there is an adequate trace (as demonstrated above for the attacks of the current stimuli).

As a result of both trace and context coding, attacks should have little influence on comparisons of timbres given long stimuli and/or ISI. It then should not be surprising that eliminating attacks had little influence on similarity judgments in Experiment 3, and the normalization task of Experiment 4. In summary, attack functions can easily be argued to be more salient features of timbre in shorter stimuli (as opposed to those used in the current experiments), where attacks are better represented in (trace) memory.

Identification of the physical properties relevant to normalization may provide important implications in understanding how the perceptual system processes auditory information. In a commentary on talker normalization, Pisoni (1990) indicated that the inability to identify the nature of source variability has hindered researchers from making significant advances in solving the problem of mapping invariant attributes of the physical signal onto abstract linguistic units. Experiments 3 and 4 address this concern in two ways for music stimuli. First, possible sources of variability were identified through stimulus alteration whose perceptual relevance was evaluated using similarity scaling procedures. Experiment 4 then provided evidence that the overtones were the timbral components most subject to normalization. Second, a possible relationship between perceived stimulus similarity and reaction time was obtained that indicated additional time was required to process signals that were judged to be more dissimilar. This increase in reaction time for dissimilar items seems to reflect the degree of adjustment that is necessary for the system to "correct" for inappropriate expectancies. We do realize that the physical alterations performed on our stimuli were rather extreme, and it thus is possible that there may be other, more subtle sources of spectral variability contributing to timbre and normalization, such as detailed aspects of upper partials (e.g., intensity patterns, decay properties).

Conclusions

In addition to providing further insight about normalization, the present study has important implications in the auditory attention domain. One new way to characterize normalization is as a manipulation of a listener's attention to stimulus features. When attending to inappropriate stimulus properties, the listener must redirect attention to appropriate settings before the relevant processing can occur. The present findings of active normalization are consistent with selective attention processes, where the perceptual system is able to set up to receive and process certain expected stimulus properties.

The present investigation has shown that normalization can be used as an important tool in identifying and defining critical auditory features utilized in signal perception. Experiment 4 demonstrated a high correlation between judged similarity and the critical parameters used in processing music timbre. Future research in normalization, whether for music or for speech, should not be limited to only demonstrating different types of normalization, but instead should focus on determining bases of listener expectations and their relations to physical characteristics of the signal. The nature of normalization then will be better established, as will a reason for why the human auditory system sometimes cannot ignore certain task-irrelevant properties of the signal.

References

Allard, F. (1976). Physical and name codes in auditory memory. Quarterly Journal of Experimental Psychology, 28, 475-482.

American Standards Association (1960). American Standard Acoustical Terminology. New York.

Beal, A.L. (1985). The skill of recognizing musical structures. Memory & Cognition, 13, 405-412.

Crowder, R.G. (1989). Imagery for musical timbre. Journal of Experimental Psychology: Human Perception and Performance, 15, 472-478.

Darwin, C.J., Turvey, M.T., & Crowder, R.G. (1972). An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. Cognitive Psychology, 3, 255- 267.

Fletcher, N.H. (1991). The Physics of Musical Instruments. New York: Springer-Verlag.

Garner, W.R. (1974). The Processing of Information and Structure. NY: Wiley.

Goldinger, S.D. (1992). Words and voices: implicit and explicit memory for spoken words. Research on Speech Perception: Progress Report No. 7 (Indiana University), 1-128.

Grey, J.M. (1977). Multidimensional perceptual scaling of musical timbres. Journal of the Acoustical Society of America, 61, 1493-1500.

Grey, J.M., & Moorer, J.A. (1977). Perceptual evaluations of synthesized musical instrument tones. Journal of the Acoustical Society of America, 62, 454-462.

Hall, M.D., & Pastore, R.E. (1992). Musical duplex perception: perception of figurally good chords with subliminal distinguishing tones. Journal of Experimental Psychology: Human Perception and Performance, 18, 752-762.

Hall, M.D., & Pastore, R.E. (1993). An auditory analogue to feature integration. Published Program for the 34th Annual Meeting of the Psychonomic Society, 16 (Abstract #174).

Handel, S. (1989). Listening. Cambridge: MIT Press.

Johnson, K. (1988). Intonational context and F0 normalization. Research on Speech Perception: Progress Report No. 14 (Indiana University), 81-108.

Jusczyk, P.W., Pisoni, D.B., & Mullennix, J.W. (1989). Effects of talker variability on speech perception by 2-month old infants. Research on Speech Perception: Progress Report No. 15 (Indiana University), 133-161.

Krumhansl, C.L., & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. Journal of Experimental Psychology: Human Perception & Performance, 18, 739-751.

Liberman, A.M., & Mattingly, I.G. (1985). Motor theory of speech perception revisited. Cognition, 21, 1-36.

Liberman, A.M., & Mattingly, I.G. (1989). A specialization for speech perception. Science, 243, 489-494.

Logan, R.J. (1990). Talker normalization and speaker recognition by humans: one mechanism or two? Unpublished doctoral dissertation. SUNY-Binghamton, Binghamton, N.Y.

Macmillan, N.A., Braida, L.D., & Goldberg, R.F. (1987). Central and peripheral processes in the perception of speech and nonspeech sounds. In M.E.H. Schouten (Ed.), The Psychophysics of Speech Perception (pp. 28-45). Dordecht, the Netherlands: Martinus Nijhoff.

Mullennix, J.W., & Pisoni, D.B. (1989). Detailing the nature of talker variability effects in speech perception. Journal of the Acoustical Society of America, 85, YY14.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (1989). Some effects of talker variability on spoken word recognition. Journal of the Acoustical Society of America, 85, 365-378.

Nusbaum, H.C., & Morin, T.M. (1989). Perceptual normalization of talker differences. Journal of the Acoustical Society of America, 85, S125.

Pastore, R.E., & Scheirer, C.J. (1974). Signal detection theory: Considerations for general applications. Psychological Bulletin, 81, 945-958.

Pisoni, D.B. (1990). Comments on talker normalization in speech perception. Research on Speech Perception: Progress Report No. 16 (Indiana University), 413-422.

Pisoni, D.B., Carrell, T.D., & Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. Perception & Psychophysics, 34, 314-322.

Pitt, M.A. (under review). Individual differences in the perception of pitch and timbre. Perception & Psychophysics.

Pitt, M.A., & Crowder, R.G. (1992). The role of spectral and dynamic cues in imagery for musical timbre. Journal of Experimental Psychology: Human Perception & Performance, 18, 728-738.

Plomp, R. (1976). Aspects of tone sensation. New York: Academic Press.

Posner, M.I. (1980). Orienting of attention. Quarterly Journal of Experimental Psychology, 32, 3-25.

Posner, M.I., Snyder, C.R., & Davidson, B.J. (1980). Attention detection signals. Journal of Experimental Psychology: General, 2, 160-174.

Saldanha, E.L., & Corso, J.F. (1964). Timbre cues and the identification of musical instruments. Journal of the Acoustical Society of America, 36, 2021-2026.

Summerfield, A.Q., & Haggard, M.P. (1975). Vocal tract normalization as demonstrated by reaction times. In G. Fant and M.A.A. Tatham (Eds.) Auditory Analysis and Perception of Speech. London: Academic Press.

Verbrugge, R.R., Strange, W., Shankweiler, D.P., & Edman, T.R. (1976). What information enables a listener to map a talker's vowel space? Journal of the Acoustical Society of America, 60, 198-212.

Whalen, D., & Liberman, A.M. (1987). Speech perception takes precedence over nonspeech perception. Science, 237, 169-171.

Wolport, R.S. (1990). Recognition of melody, harmonic accompaniment, and instrumentation: Musicians vs. nonmusicians. Music Perception, 8, 95-106.

Acknowledgments

This research was supported by grant F496209310033 from the Air Force Office of Scientific Research and grant BNS8911456 from the National Science Foundation. Opinions, findings, conclusions, and recommendations are the authors' and do not necessarily reflect views of the granting agencies.

Footnotes

1. Because the octave location on the synthesizer corresponded to a different frequency range for the woodwinds, the chords produced by the woodwinds were lower in frequency (approximately one octave) compared to the other instruments.

2. Context coding should be richer given extensive experience with stimuli. Thus, listeners with greater music experience could be conjectured to additionally have adequate encoding of more subtle stimulus properties, like the attack functions of our stimuli.

Table 1. Mean percent correct and mean RT (plus standard error) for same and different chord trials in Experiment 1 for single versus mixed instrument conditions. Results for different chord trials are further partitioned by the number of notes which differ between compared chords.

	SINGLE INSTRUMENT				MIXED INSTRUMENT			
Chords	Accuracy		RT in ms		Accuracy		RT in ms	
SAME	98.2	(0.9)	1069	(36)	60.9	(1.6)	1175	(34)
DIFFERENT	88.5	(3.3)	1210	(50)	66.6	(1.4)	1204	(55)
1-note:	80.7	(3.1)	1146	(30)	62.3	(2.0)	1199	(23)
2-note:	95.3	(1.6)	1113	(37)	70.1	(3.3)	1199	(32)
3-note:	98.0	(1.2)	1060	(47)	74.3	(3.5)	1230	(55)

Table 2. d' and RT for each instrument combination across all AX chord discrimination trials (independent of note differential for different trials) in Experiment 1, including overall means and standard errors.

Instrument Effects on d' and RT (Experiment 1)

Instrument A Chord	Instrument X Chord	d'	RT (Same Chord)	RT (Diff.Chord)
Piano	Piano	4.25	1021	1087
	Brass	1.16	1172	1322
	Woodwind	0.60	1433	1270
	String	1.55	1190	1299
	Harpsichord	0.95	1238	1220
Brass	Piano	1.81	1173	1243
	Brass	4.64	1088	1155
	Woodwind	0.99	1472	1229
	String	1.53	1185	1228
	Harpsichord	1.11	1243	1172
Woodwind	Piano	0.67	1193	1207
	Brass	0.50	1255	1301
	Woodwind	4.06	1090	1143
	String	0.47	1318	1325
	Harpsichord	1.04	1200	1333
String	Piano	1.44	1104	1214
	Brass	1.46	1238	1279
	Woodwind	0.69	1401	1263
	String	4.37	1129	1147
	Harpsichord	1.13	1328	1212
Harpsichord	Piano	1.26	1198	1181
	Brass	0.60	1284	1365
	Woodwind	0.89	1202	1334
	String	0.87	1279	1345
	Harpsichord	3.81	1040	1070
	Mean	1.67	1219	1237
	s.e.	0.29	54	59

Table 3. RT data on same chord trials for Subjects 7 and 13 (Experiment 1).

Subject 7 Chord	Same Instrument	Different Instrument
C-Major	853	946
C-Minor	676	964
E♭-Major	761	1007
E♭-Minor	797	989
Mean	**903.3**	**1040.5**
St. Dev	**64.3**	**23.3**
Subject 13		
C-Major	953	1127
C-Minor	1109	1098
E♭-Major	1135	1088
E♭-Minor	942	1105
Mean	**1034.8**	**1104.5**
St. Dev.	**87.8**	**14.3**
Subject 2		
C-Major	840	983
C-Minor	1015	1054
E♭-Major	987	922
E♭-Minor	888	971
Mean	**932.5**	**982.5**
St. Dev.	**96.6**	**54.4**
Subject 21		
C-Major	1032	1100
C-Minor	1011	1066
E♭-Major	980	1054
E♭-Minor	962	1075
Mean	**996.2**	**1073.8**
St. Dev.	**31.3**	**19.5**

Table 4. Mean RT and standard error data across subjects for each cue-instrument combination for Experiment 2.

Cue	Instrument	Mean RT	St. Error	Cue	Instrument	Mean RT	St. Error
VALID TRIALS:				INVALID TRIALS:			
P	P	642.99	83.17		P	B	1120.88	167.38
B	B	767.36	92.72		P	S	981.71	136.27
S	S	758.19	85.83		P	H	802.52	73.30
H	H	744.66	81.32		B	P	885.24	147.29
		Mean 728.30			B	S	1020.08	139.24
					B	H	1029.34	167.23
NEUTRAL TRIALS:					S	P	953.77	118.35
X	P	732.67	116.86		S	B	1153.19	190.01
X	B	849.62	91.89		S	H	1044.03	131.04
X	S	821.63	98.56		H	P	712.62	89.19
X	H	755.14	67.96		H	B	950.21	128.87
		Mean 789.76			H	S	1021.16	177.24
							Mean 972.89	

Table 5. Decoding of symbols used in Figures 2-4.

Symbol	Stimulus	Symbol	Stimulus
Ps	Synthetic Piano	Pn	Natural Piano
Bs	Synthetic Brass	Tn	Natural Trumpet
Ws	Synthetic Woodwind	Fn	Natural Flute
Ss	Synthetic Strings	Sn	Natural Strings
Hs	Synthetic Harpsichord	Hn	Natural Harpsichord
Pc	Cut-Attack Piano	Pf	Filtered Piano
Tc	Cut-Attack Trumpet	Tf	Filtered Trumpet
Fc	Cut-Attack Flute	Ff	Filtered Flute
Sc	Cut-Attack Strings	Sf	Filtered Strings
Hc	Cut-Attack Harpsichord	Hf	Filtered Harpsichord

Table 6. Mean similarity ratings for synthetic versus natural instrument comparisons.

Instrument Comparison	Mean Rating	St. Error
Sn - Ss	6.51	0.14
Pn - Ps	6.64	0.18
Hn - Hs	5.76	0.31
Fn - Ws	5.60	0.34
Tn - Bs	6.31	0.18

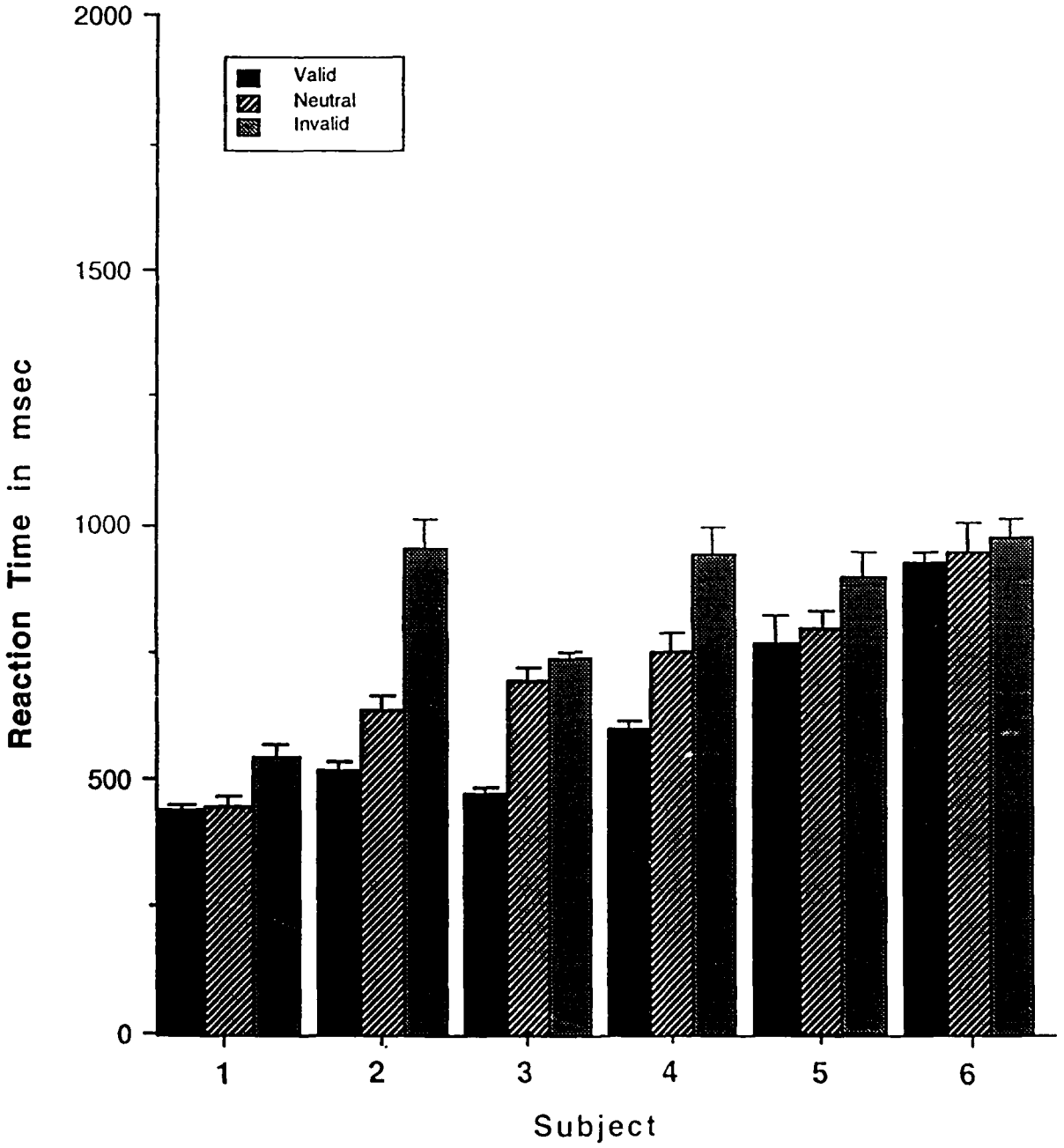
Figure Captions

Figure 1. Mean reaction times for individual subjects for valid, neutral, and invalid visual, instrument cue trials in Experiment 2.

Figure 2. (a) Multidimensional similarity scaling (dimensions 1 and 2) for individual natural and synthetic instruments in Experiment 3, Condition 1; (b) Multidimensional similarity scaling (dimensions 1 and 2) for intact natural stimuli and cut-attack versions in Experiment 3, Condition 2; (c) Multidimensional similarity scaling (dimensions 1 and 2) for intact natural stimuli and filtered versions in Experiment 3, Condition 3.

Figure 3. Linear regression of mean reaction times as a function of similarity ratings in Experiment 4, including two problem conditions.

Individual Effects of Cue Validity on Reaction Time



Figure

SIMILARITY MATRIX FOR NATURAL AND SYNTHETIC INSTRUMENTS

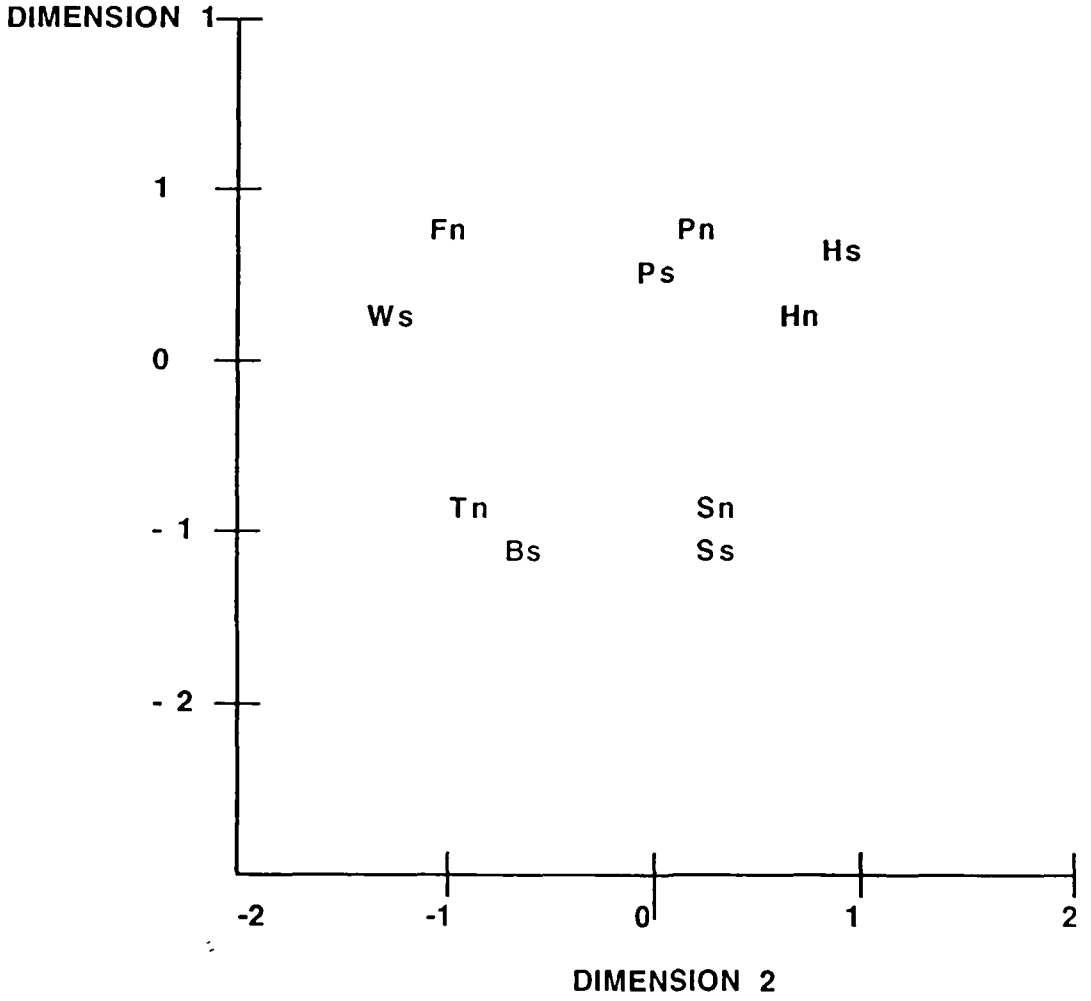


Figure 2a

Similarity Matrix For Intact and Cut-Attack Stimuli

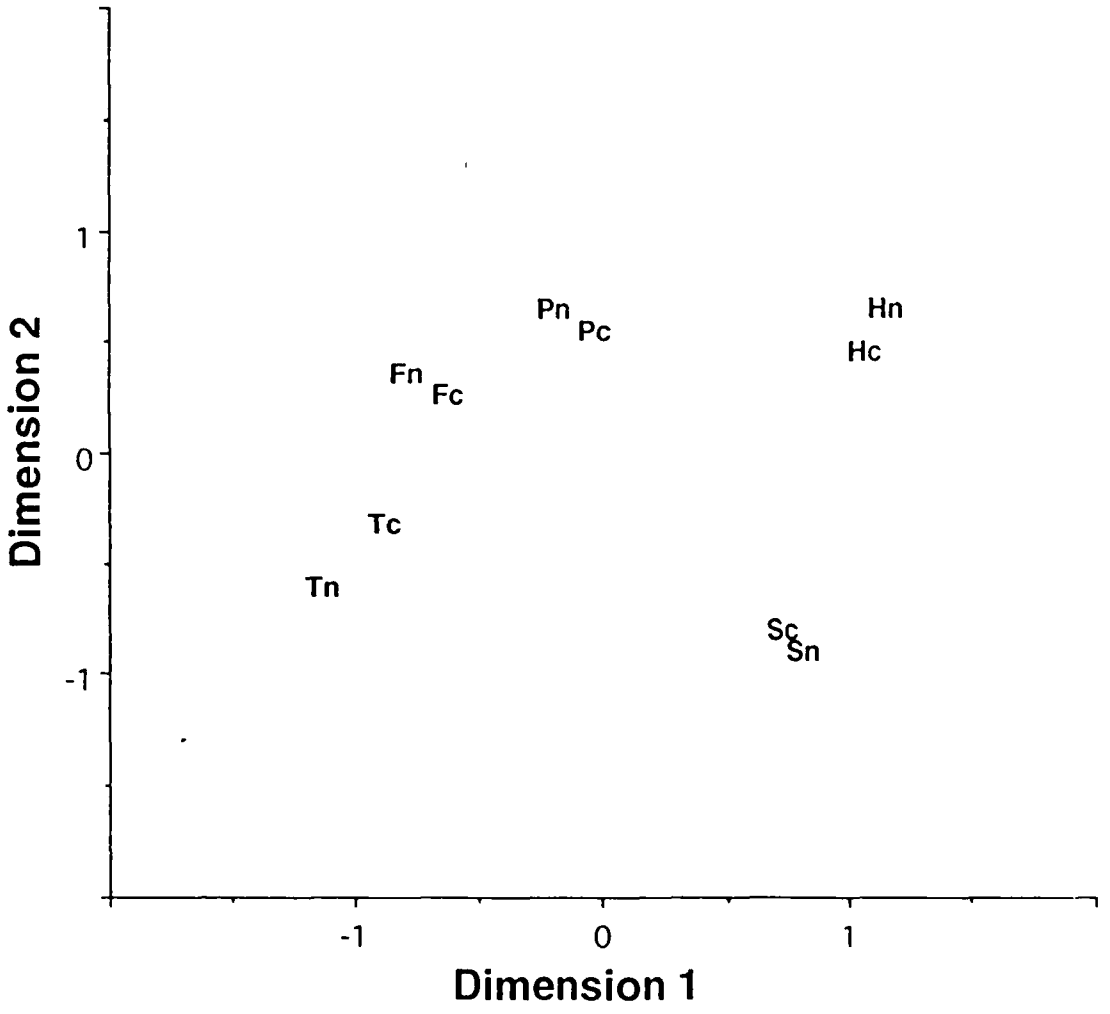


Figure 2b

Similarity Matrix For Intact and Filtered Stimuli

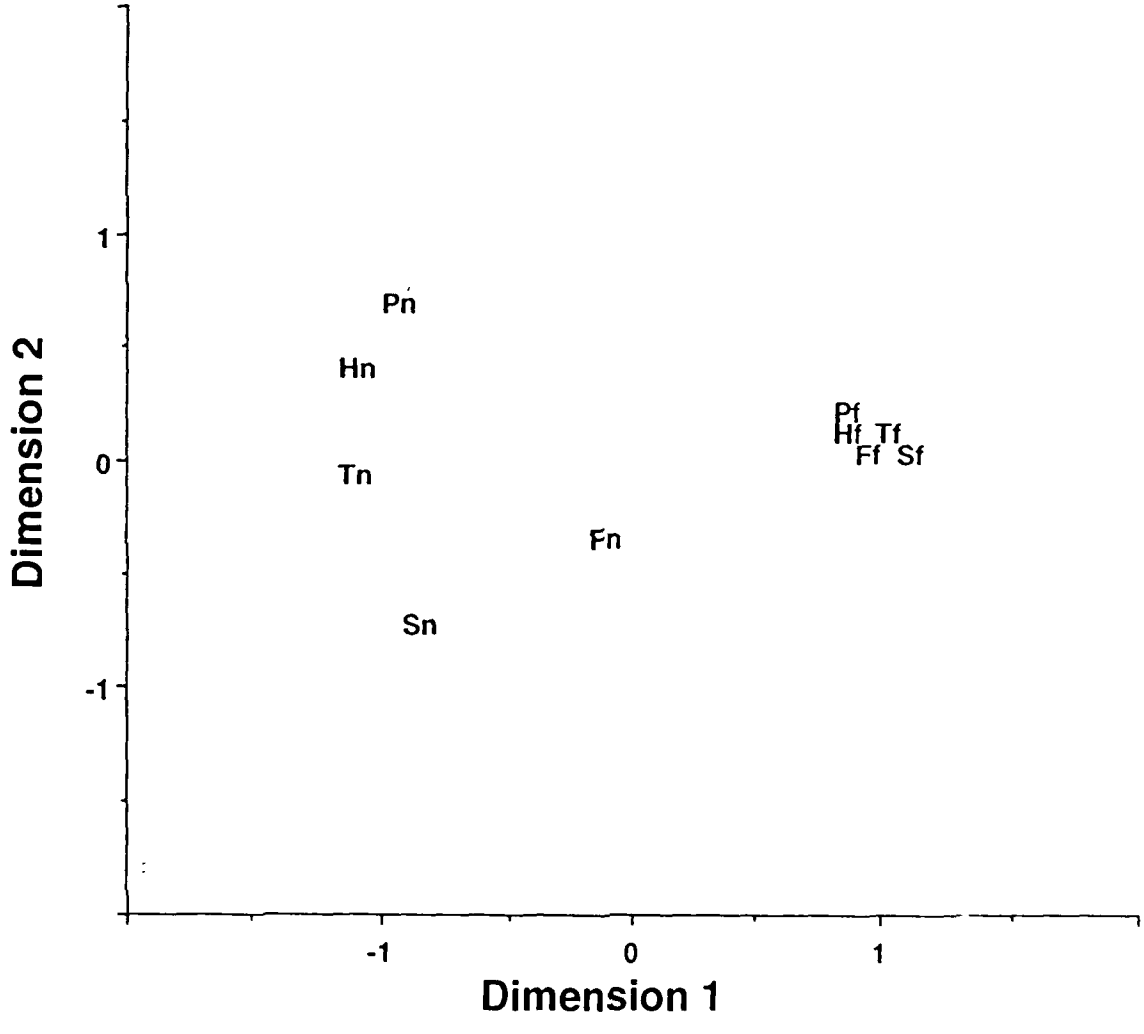
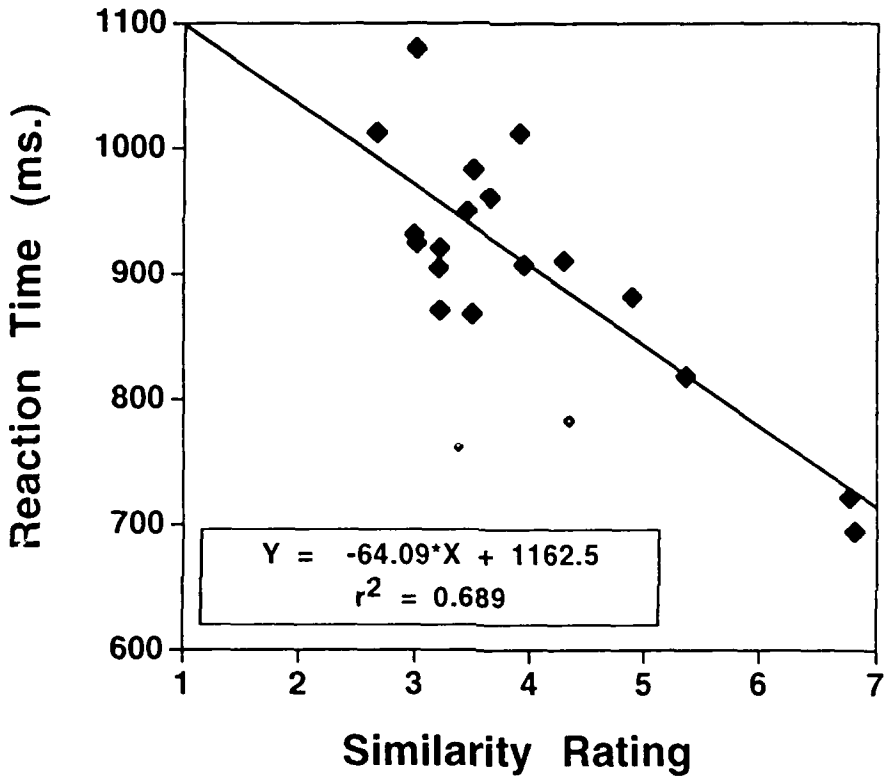


Figure 2c

Music Normalization: RT versus Similarity Rating



Figure

Effects of Stimulus Complexity on the *Perceptual Organization of Musical Tones*

Michael D. Hall and Richard E. Pastore
Center for Cognitive and Psycholinguistic Sciences
State University of New York at Binghamton
Binghamton, NY 13902-6000

Abstract

Duplex perception (DP) occurs when one stimulus or stimulus component contributes simultaneously to two distinct percepts. Two AX discrimination experiments were conducted to quantitatively evaluate the effects of one factor, stimulus complexity (by manipulating the number of frequency components common to major and minor chords), which would be predicted by Gestalt principles of perceptual organization to affect incidence of fusion in DP stimuli. Experiment 1 demonstrated frequent fusion of bases with a contralateral distinguishing tone. Data from both experiments first provide some indirect evidence against the claim made by some supporters of speech modularity that musical DP research is really demonstrating triplex perception (i.e., perception of base, tone, and chord). The experiments further revealed that when major/minor chords are presented contralateral to a different distinguishing tone, chord ear perception was altered. These alterations, which included perceptual migrations and fusion of contralateral distinguishing tones, did not depend on stimulus position within a trial and was a direct function of stimulus complexity. The results are discussed in terms of the relationship between stimulus complexity and figural goodness, and are evaluated as possible examples of stimulus dominance and illusory feature conjunctions.

One major goal of auditory perception research is to identify general principles of perceptual organization. This goal frequently has taken the form of establishing the stimulus variables critical to the perception of complex stimuli. Much of this literature has concentrated on laboratory phenomena involving stimuli which consistently give rise to ambiguous or illusory percepts [e.g., the octave illusion (Deutsch, 1974)]. One such laboratory phenomenon, duplex perception (DP), typically has been demonstrated with speech stimuli. In DP one stimulus, or stimulus component, simultaneously contributes to two distinct percepts. DP is claimed to represent a violation of the rule of disjoint allocation (derived from the Gestalt principle of belongingness), which states that one stimulus/component can or'y contribute to one perceptual stream (Bregman, 1987; Mattingly & Liberman, 1989; but see Bregman, 1990).

The following experiments with musical stimuli evaluated one factor which might affect the incidence of fusion in DP. In addition to addressing an existing skepticism about the validity of musical DP, the results of the experiments are consistent with the illusory conjunction of auditory features, and thus raise questions about the role of attention in tonal stimulus processing. Before reviewing the current experiments, a brief history of DP research will be presented, including a brief evaluation of theoretical issues DP has been used to address.

DP Phenomena and the Claim for a Phonetic Module

DP was first demonstrated by physically splitting components of synthetic versions of /da/ and /ga/ syllables which differ in place of articulation (Rand, 1974). A common form of DP for speech (DPS) entails the (dichotic) presentation of the third formant (F3) transition and the remainder of the syllable (or base) to separate ears (e.g., Mann, Madden, Russell, and Liberman, 1981). The isolated F3-transitions are perceived as chirps. The F3-transitions also distinguish between /da/ and /ga/, with bases presented in isolation often only heard as somewhat ambiguous or neutral syllables (i.e., usually not consistently labelled as either /da/ or /ga/). When presented dichotically at normal listening levels, the result is two simultaneous perceptions, with the transition playing a critical role in each: (1) perception of the transition as an isolated chirp in one ear, and (2) perception of a complete /da/ or /ga/ syllable (base plus transition) in the base ear.

DPS is cited extensively by Liberman and colleagues (e.g., Liberman, Isenberg, and Rakerd, 1981; Liberman and Mattingly, 1985, 1989a, b; Whalen and Liberman, 1987) as evidence for the existence of a specialized, biologically significant, phonetic module. This argument presumes that DPS percepts are the result of two, separate, distinct types of processing performed on the transition. Chirp perception reflects the common nonspeech operation of a general, open, auditory module where perception corresponds relatively directly to the physical properties of the signal (pitch, loudness, and timbre). Processing by the closed speech module instead results in the perception of a CV syllable where stimulus and perception do not directly coincide except in terms of phonetically relevant stimulus properties.

DP replications with analogous nonspeech stimuli, including demonstrations using musical tones (Collins, 1985; Hall and Pastore, 1992; Pastore, Schmuckler, Rosenblum, and Szczesiul, 1983) and door slamming sounds (Fowler and Rosenblum, 1990), have questioned DP-based conclusions for modularity, minimally demonstrating the operation of other auditory modules in DP which mirror processing by the speech module. We believe that such nonspeech conditions reveal that DP findings to date can be addressed equally well from modular and general auditory perspectives. Musical DP was first obtained by Pastore, et al. (1983, and later replicated, with reduced incidence, by Collins, 1985). A tone distinguishing a major from minor chord (E or E♭, which is presumed to play a role analogous to the transition in DPS) was dichotically presented with the remainder of the chord, the corresponding fifth interval (the C-G base). Many musically trained subjects reliably identified hearing both the isolated tone in the appropriate ear and a complete, fused, major or minor chord in the base ear. As with isolated transitions in DPS, isolated tones were labelled less accurately than chords perceived by integration of tones and base. Thus, the musical base

seems to act as a harmonic frame of reference upon which judgments on the distinguishing tone can more easily be made.

Findings of musical DP have been criticized by phonetic modularity supporters as not having ruled out triplex, rather than duplex, perception (Mattingly and Liberman, 1988). Triplex perception (TP) presumes that subjects accurately hear both the distinguishing tone and base at their respective physical locations, and additionally perceive the integration of these stimuli as a centrally localized chord. However, we are not aware of any empirical evidence which suggests that TP exists for any stimuli. In fact, both musical and speech DP subjects generally report hearing stimuli at only two locations. As a secondary issue, the present experiments investigate the likelihood of TP with musical stimuli.

Gestalt Notions and DP

If both speech and nonspeech DP stimuli are processed in a singular manner, then speech and nonspeech DP need not necessarily reflect the operation of distinct modules. For example, according to Gestalt terminology (e.g., Wertheimer, 1958), syllables and chords both should represent "good" (simple, organized, unified perceptions of stimuli consisting of several components that frequently occur together), "strong" (resisting analysis into separate components) figures. These good, strong figures are thus more easily perceived with less information (e.g., stimulus energy or impoverished components) than is required to separately perceive components critical to their perception (e.g., transitions or chord-distinguishing tones), resulting in "closure" (see below) and thus the apparent precedence-taking DP findings.

Bregman (1987, 1990) has suggested that both speech and nonspeech DP arise when there is sufficient conflict between cues used to segregate and to integrate two stimuli. Differential localization or quality is a major cue for segregating base and distinguishing component which could compete against integration cues which reflect either Gestalt principles of perceptual organization (e.g., good continuation and frequency/temporal proximity) or stimulus-specific, schema-based properties. Fusion to perceive chords or syllables then would presumably result from (1) the synchronous presentation of components, (2) the end frequency of the transition corresponding to the initial steady state frequency in the base in DPS, and (3) the tone and fifth maintaining simple integer frequency ratios in musical DP. Motivated by such an analysis, Ciocca and Bregman (1989) have demonstrated the weakening of DP for speech syllables when the contralateral distinguishing component (transition) is part of a coherent stream reflecting the principles of similarity and good continuation. The Gestalt conceptualization should not only encompass both speech and nonspeech auditory examples, but also has been suggested by Bregman as generalizing across modalities.

Motivation for Current Research

DP research cannot at present (and probably can never) provide unequivocal evidence in support of a postulated phonetic (and/or musical) module. In focusing on modularity, previous DP research has often overlooked critical perceptual issues which the DP paradigm is well suited to address. Perception of stimuli used in studies of DP can provide a unique opportunity to evaluate the contribution of specific variables to auditory perceptual organization, by (1) revealing the conditions necessary for perceptual integration, (2) evaluating the relative saliency of organizational cues, and (3) specifying the nature of attentional and perceptual limits of the auditory system.

The strength of general perceptual (e.g., Gestalt) explanations of processes underlying DP can be evaluated by revealing the various conditions necessary for frequent fusion with both speech and nonspeech stimuli. In so doing, we will gain a better understanding of (1) the critical cues for integration and segregation of stimuli, and (2) how these cues operate in the presence of other consistent or conflicting sources of information for grouping stimuli. After the separate conditions for fusion in speech and nonspeech DP are established, analogous speech and nonspeech conditions might be investigated in a more realistic attempt to resolve the phonetic modularity issue for DP research. Then, if analogous stimuli always provide similar patterns of perception, DP might reflect the operation of general auditory principles. If, however, significantly different perceptual tendencies are obtained for analogous speech and nonspeech stimuli, the nature of distinct modes of processing could begin to be established.

Some stimulus variables critical to the incidence of DPS have been determined. Characterizing DP as a form of spectral/ temporal fusion, Cutting (1976) evaluated the effects of several variables on fusion of speech stimuli. Fusion in DPS was relatively insensitive to changes in intensity (also see Whalen & Liberman, 1987) and frequency, but diminished with increasing asynchrony of component stimuli. Similar evaluations of possibly critical stimulus variables for musical DP have been lacking.[1] The present experiments represent an initial attempt to evaluate stimulus factors which may be critical to fusion in musical DP. Base complexity (defined as the number of invariant base components shared between at least two labelling categories) was selected as one variable which could affect the incidence of fusion. For the present purposes, increasing musical base complexity will be defined as adding tones of different chroma to the original C-G (fifth) base.

Generally, western music listeners have not been presented the base in isolation. The base also cannot be resolved as a major or minor chord, which represent more commonly heard chord structures. Therefore, because of the unusual nature of the base, it is argued that the bases used in the current experiments all represent open forms. Manipulating the complexity of the base should alter the figural goodness of the chord resulting from the fusion of distinguishing tone and base (see below). As a result, we will use the term "base complexity" not only to describe our dichotic stimulus manipulation, but also to refer to alterations in stimulus (chord) complexity.

Base complexity represents one factor which long-standing Gestalt principles of perceptual organization would predict to influence the rate of perceptual integration. The Gestalt principle of closure is defined as the perceptual tendency to complete (close) physically incomplete (open) forms, resulting in the perception of good, strong figures instead of poor, weak figures. Component stimuli within a closed form are perceived as belonging to a more stable representation than if separately perceived (Boldt, 1942). Assuming that the base in DP is a relatively open form, stimulus properties which increase the goodness and strength of a (major or minor) chord could be regarded as increasing the tendency toward closure based upon the fusion of tone

and base. In other words, dichotic configurations which are more readily fused and thus closed (or less open) could be assumed to represent better articulated, more stable forms. Our major interest in determining the likelihood of any given perceptual organization (e.g., fusion) as a function of complexity was that such a determination may reveal the nature of the relationship between stimulus complexity and figural goodness for the current musical stimuli.

Does increasing base complexity of musical stimuli result in more open or more closed stimuli? The Gestalt principles of good continuation and (both frequency and temporal) proximity suggest that chords with many tones may represent stronger, more figurally good forms than chords consisting of fewer tones (Wertheimer, 1958). Given the perceptual tendency to perceive good, strong figures, distinguishing tones then should be more easily integrated with bases of greater complexity. The tendency to fuse a tone and base in the base ear then should increase as the number of tones in the base increases.

Alternatively, because the frequency ratios between component tones cease to be in simple integer ratios (Dowling and Harwood, 1986), increasing the number of tones decreases the overall consonance of chords. As a consequence, increasing complexity instead may decrease figural goodness for base stimuli and bases fused with distinguishing tones. Thus, fusion of tone and base may become less likely with the addition of more tones to the base. Therefore, we cannot make explicit predictions regarding the probability of fusion with varied base complexity, but rather leave this issue as an important perceptual question for which the present research can provide an empirical answer.[2]

EXPERIMENT 1: Establishing Effects of Base Complexity

Experiment 1 estimated the probability of many distinct perceptual organizations as a function of changes in stimulus (base) complexity. These organizations included not only the probability of fusion and TP, but also perceptual configurations common to other (speech and visual) stimuli.

Method

Some of the current methods were motivated by DP findings. Although used in previous musical DP demonstrations (Pastore, et al., 1983; Collins, 1985), major/minor chord labelling performance is not equally good across subjects, even when the subjects are musicians. This lack of consistent performance across subjects complicates the understanding of whether or not fusion is in evidence for some subjects. Most people without extensive musical backgrounds, however, can reliably perceive differences between major and minor chords, even if some of these individuals cannot consistently apply appropriate labels to each chord. Since ability to discriminate chords produced solely by fusion is strong evidence for fusion, an AX procedure was used.

Subjects. All subjects had studied at least one musical instrument (although not always an instrument capable of producing chords) and thus, in theory, understood the distinction between major and minor chords. However, because we accurately expected that musical expertise of possible subjects would vary widely, an a priori performance criterion of better than chance performance for binaural chord discrimination was adopted in both experiments to insure that all subjects had a working understanding of the perceptual difference between major and minor chords. Since failure to accurately discriminate binaural chords made evaluation of dichotic perceptual organization impossible, we discarded the data from any subject who did not meet the a priori criterion. In Experiment 1, 11 SUNY-Binghamton undergraduates who met the a priori performance criterion participated as subjects in partial fulfillment of course requirements.[3] The experiment lasted approximately 45 minutes.

Materials. Stimuli were generated from digitally sampled piano tones (Yamaha AWM Sound Expander EMT-10), recorded on cassette tape, and digitized (12-bit, 10 kHz sample rate) for on-line computer presentations (with 4 kHz antialiasing filter). All tones were of equal length (1424 ms), and were from an equi-tempered interval scale with the following tone chroma and frequency in Hz: C (266), E♭ (316), E (335), G (398), A (447), B (501), and D (597). Tones then were digitally mixed to produce the various base complexes and chords. All stimuli were presented over TDH-49 headphones at 75 dB SPL peak amplitude.

E and E♭ tones always distinguished chords as major (e.g., C-E-G) or minor (C-E♭-G). Chords were additionally distinguished by the number of tones (2, 3, or 4) constituting the base in dichotic trials; the 2-, 3-, and 4-tone bases were C-G, C-G-A, and C-G-B-D, respectively.

Procedure: Binaural Discrimination. Upon consent, subjects ran a block of 80 randomized binaural same-different (AX) discrimination trials intended as a baseline measure of subject performance for subsequent dichotic trials derived from the same stimuli.[4] Both binaural and dichotic trials consisted of the A (standard) and X (comparison) stimuli separated by a 1500 ms ISI, and ended with a 2 s response interval.

Each of 6 possible chords was presented as the A stimulus on 10 trials. The chords were: C-major (C-E-G), C-minor (C-E♭-G), C-major 6th (C-E-G-A), C-minor 6th (C-E♭-G-A), C-major 9th (C-E-G-B-D), and C-minor 9th (C-E♭-G-B-D). Each A stimulus (e.g., a major or minor C 6th chord) was paired equally often with itself and its alternative (minor or major 6th) chord as X stimuli. The remaining (20) trials were designed to insure that subjects also could reliably distinguish isolated E and E♭ tones, presenting each tone as the A and the X stimulus with an independent probability of 0.5. These binaural conditions also provided the empirical qualification criterion for subjects through a direct quantification of "same"/"different" response tendencies for each condition and level of base complexity. These results later will be used to correct estimated perceptual probabilities for response tendencies.

Dichotic Discrimination: Stimuli. After a short break, subjects began a block of 240 dichotic trials. Subjects were instructed to label each AX stimulus pair as same or different only with respect to the ear in which bases or complete chords were presented (henceforth, the target ear) and to ignore the information presented to the other ear (henceforth, the contralateral ear). Target ear assignment remained constant for a given subject, but was counterbalanced across subjects.

Table 1 summarizes the dichotic stimuli along with possible nonveridical perceptual organizations (i.e., perceptions

differing from the physical configuration of tones) for each stimulus used. In summarizing the conditions throughout both tables and text, only 2-tone base components will be presented. Stimuli for 3- and 4-tone bases can be obtained by respectively adding A and B-D tones to the C-G base. Target ear (t.e.) tones will always be displayed to the left of double vertical lines, with contralateral ear (c.e.) tones presented to the right.

Insert Table 1

The nonveridical percepts critical to our original hypotheses are instances of fusion (displayed in column 2 of Table 1). Fusion arises when subjects integrate a contralateral distinguishing tone with target ear information to perceive a single, unified percept in the target ear. For demonstration purposes, fusions are displayed as instances of DP, with the simultaneous perception of the distinguishing tone as a separate event. If chords with many tones represent better-articulated figures, then fusion should occur more frequently as base complexity increases when presented the base and a distinguishing tone to contralateral ears. If increasing complexity instead decreases figural goodness, then fusion of tone and base should decrease with increasing complexity. When two distinguishing tones are presented dichotically (one physically mixed with the base), fusion should result in the perception of a chord (consisting of both E and E♭ tones) that is neither major nor minor, which should be an unstable (dissonant) figure. Thus, such fusions should not occur frequently. Additionally, if increasing complexity increases figural goodness, fusion of contralateral distinguishing tones should decrease as a function of increasing complexity. Conversely, if increasing complexity decreases figural goodness, fusion of distinguishing tones should increase as a function of increasing complexity.

Another plausible, nonveridical percept (shown in column 3 of Table 1) is "migration". Migrations were proposed to occur only when subjects were contralaterally presented different chord-distinguishing tones (E and E♭, one of which was mixed with the base). In these instances, both distinguishing tones would be perceived, but contralateral to their physical locations. While intuitively improbable, migrations, like fusions, maintain the unified perception of a figurally strong chord in the target ear and a contralateral chord-distinguishing tone. Such invalid assignment of feature locations has been demonstrated in the visual attention literature (e.g., Treisman, 1990), and should be likely to occur in audition; we will later describe how notions from the visual domain might apply to musical stimuli.

Column 4 of Table 1 depicts another nonveridical percept which has been demonstrated with speech stimuli. Repp (1978a, b, and c) identified which of several dichotic pairs of CV syllables fused perfectly, giving rise to the perception of a single syllable. Labelling of these stimuli often exhibited patterns of stimulus dominance, a ". . . tendency of one stimulus in a specific dichotic pair to receive more correct responses than the other stimulus, regardless of the ear in which it occurs (p. 133)." A dominant stimulus element therefore contributes to perception regardless of where it is presented.

Stimulus dominance could occur only when two dichotic distinguishing tones were simultaneously presented, one of which was physically mixed with the base (e.g., C-E-G ‖ E♭). Dominance would result in the perception of the isolated distinguishing tone in both ears, preventing perception of the other distinguishing tone (i.e., C-E♭-G ‖ E♭). While responses for any singular condition cannot distinguish between migration or stimulus dominance, comparisons across conditions (addressed in the general discussion) will allow an evaluation of the likelihood of both percepts.

Possible triplex percepts are shown in column 5 of Table 1. Due to a lack of any evidence for TP, TP was hypothesized not to occur. If TP occurs in a manner consistent with the postulations of Liberman and colleagues (e.g., Mattingly and Liberman, 1989), subject responses should be equivalent to those given veridical target ear perception. If, on the other hand, subjects ignore the instructions to respond on the basis of target ear perception, and, instead, are "distracted" to respond to the triplex percept at a central, abstract position, then responses based on TP would be indistinguishable from instances of fusion. Therefore, the incidence of TP cannot exceed the joint probability of veridical perception and distraction. Although it is impossible to directly evaluate distraction to a postulated triplex percept, we did test for "distraction" to respond to contralateral ear information in Experiment 2. If such distractions are rare, then responses reflecting an integration of tone and target ear information (base or chord) would be more consistent with fusion than TP.

Dichotic Conditions. All conditions were generated using the stimuli summarized in column 1 of Table 1, but with varying levels of base complexity. The various conditions were designed to (1) assess subject tendencies to fuse both, one, or neither (A, X) stimuli as a function of component arrangements and levels of complexity, and (2) to minimize possible overall response biases. In each of four conditions 60 randomly mixed trials were presented, with 20 trials for each level of base complexity. Ten trials presented one configuration of distinguishing tone(s) and base; the other 10 trials substituted E for E♭ and vice versa.

A listing of dichotic conditions for both experiments, including responses expected for each possible perceptual organization, is provided in Table 2. Only one configuration of distinguishing tones is listed; the alternative configuration can be obtained by substituting E for E♭ and vice versa. Conditions are discussed in the order in which they appear in the table so that the table may be used as a reference throughout description of conditions and response predictions.

Insert Table 2

Each condition was designed to evaluate the perceptual organization indicated by the condition label. For example, the FUSE-EITHER Condition is an exclusive (XOR) condition for fusion, where "different" responses will result from fusion of either the A or X stimulus, but not both. All other percepts would result in "same" responses. Since the physical configuration of the A and X stimuli were identical in this condition, "same" responses were hypothesized to predominate

The FUSE-NEITHER Condition was generated by substituting the alternative distinguishing tone in the X stimulus of the FUSE-EITHER Condition. Only by appropriately perceiving target ear information in both stimuli would subjects produce "same" responses; fusing either (or both) distinguishing tone(s) would instead result in nonequivalent percepts. Given the demonstrated tendency to fuse these musical stimuli (e.g., Pastore, et al., 1983), high rates of "different" responses were hypothesized. If chords consisting of many tones represent better-articulated figures, we would predict an increased frequency of "different" responses with increasing base complexity. Conversely, if chords with many tones represent poorly articulated figures, "different" responses should decrease with increasing base complexity.

The remaining conditions involved dichotic presentation of both distinguishing tones (one physically mixed with the base), and thus allowed evaluation of migration or dominance. The FUSE-1 Condition was designed to determine the probability of fusing only the X stimulus. "Same" responses only would be obtained when subjects appropriately perceived target ear information in the A stimulus and fused the X stimulus (both resulting in the target ear perception of C-E-G in the example). Due to the expected high rate of fusion of the X stimulus (similar to the FUSE-NEITHER stimuli above), predominantly "same" responses were hypothesized. Because fusion of contralateral distinguishing tones (or TP) would result in the perception of a dissonant, unstable form, and migration would require mislocating two component tones, a low probability of altered target ear perception was expected for the A stimulus. Furthermore, if increasing complexity increases figural goodness "same" responses should increase as a function of increasing complexity (and, conversely, if increasing complexity decreases goodness "same" responses should decrease).

Finally, the FUSE-BOTH Condition could reflect the likelihood of fusing contralateral distinguishing tones of both stimuli to perceive figurally poor, weak chords (e.g., C-E♭-E-G), which would result in "same" responses. As noted for the FUSE-1 Condition, neither fusion nor migration of these stimuli was expected. However, migration or dominance of distinguishing tones for either the A or X stimulus also would result in "same" responses. Other perceptions would result in "different" responses, including the hypothesized veridical perception.

Results and Discussion

Binaural (Baseline) Discrimination. Mean accuracy in terms of percent correct (with standard errors) on binaural discrimination trials is shown in the top panel of Table 3. All subjects discriminated isolated tones (E and E♭) with perfect accuracy, and discriminated chords at high levels of accuracy. If we assume that dichotic fusion provides the same underlying basis for perception as presentation of physically mixed stimuli, then mean accuracy rates for binaural discrimination should provide a baseline of chord discriminability for evaluating dichotic results.[4]

Chord discrimination results were analyzed in a 2 X 3 ANOVA, with chord (same and different trials) and complexity (number of tones) as the respective factors. The only significant effect was the chord X complexity interaction ($F[2,20]=5.344$, $p=.0138$). The nature of the interaction can be seen in the table of means. Accuracy slightly increased with increasing complexity for same chord trials (revealed by a nonsignificant simple main effect, $F[2,20]=1.477$, $p=.252$). However, accuracy significantly decreased on different chord trials with increasing complexity (simple effect, $F[2,20]=4.045$, $p=.033$). This interaction suggests that with increasing chord complexity, there may be an increase in overall perceived similarity between major and minor chords, with the processing of chord distinguishing tones becoming more difficult. Later, these binaural results were later used to adjust mean performance on dichotic trials (below) to assess the probability of fusion in each condition.

Dichotic Discrimination: Overall. Mean percentage of "same" responses for each dichotic condition and level of base complexity are shown in the lower panel of Table 3. An initial 4 X 3 ANOVA was conducted with dichotic condition and base complexity as respective factors. As expected, all effects were significant, including the main effects of base complexity ($F[2,20]=23.269$, $p<.0001$) and condition ($F[3,30]=26.144$, $p<.0001$), plus their interaction ($F[6,60]=7.96$, $p<.0001$). Effects within conditions are included below in the context of individually evaluating the perception of contralateral tone and base, as well as contralateral tone and chord.

Insert Table 3

Because the combination of perceptual organizations leading to a response were different across conditions, it was possible to solve for the probability of particular organizations using a set of simultaneous probability equations based on data across conditions. Therefore, where applicable, the dichotic means from Table 3 also were submitted to a series of probability formulae. These formulae also used the binaural chord discrimination results to adjust for the tendency to perceive chords more similarly with increasing complexity.

The formulae were based on a few reasonable simplifying assumptions, the validity of which was verified for similar stimuli in an earlier musical DP study (Hall and Pastore, 1992). First, incidence of a given type of perceptual organization was assumed not to vary for isolated E and E♭ distinguishing tones. Thus, results were collapsed with respect to distinguishing tones within each (stimulus and) condition. Second, since there is no basis for comparison at the time the A stimulus is presented, perception of the A stimulus should not depend on the nature of the X stimulus. Thus, incidence of a given type of perceptual organization was assumed to be equivalent across conditions for A stimuli of similar structure. Finally, for conditions based solely upon stimuli containing a contralateral distinguishing tone and base (i.e., FUSE-EITHER and FUSE-NEITHER Conditions), equal, independent probabilities of fusion were assumed for A and X stimuli. The results of probability formulae for both experiments are shown in Table 4, calculated individually for each type of stimulus and each level of base complexity. Derivations of formulae are found in the Appendix.

Insert Table 4

Perception of Contralateral Tone and Base. Simple effects of base complexity were not significant for the FUSE-EITHER and FUSE-NEITHER Conditions ($F=.018$, $p<.900$, and $F=2.144$, $p<.143$, respectively). As a reminder, the high incidence of "same" responses on FUSE-EITHER trials could reflect either fusion or veridical perception of both A and X stimuli, or both. The moderate rate of "different" responses on FUSE-NEITHER trials was consistent with the fusion of tone and base in one or both stimuli.

Means from both conditions were submitted to probability formulae to estimate the incidence of fusion and veridical perception of contralateral tone and base. The probability of fusing both stimuli in the FUSE-EITHER and FUSE-NEITHER Conditions (calculated using Appendix Eq. 3'), displayed in the first panel of Table 4, was moderately high, with a slight decrease from 2- to 3-tone levels of base complexity. Assuming equal, independent probabilities for fusing A and X stimuli, we obtain the probability of fusing either stimulus (see Table 4), which occurred at a substantial rate ($p=0.65$ to 0.75).

Perception of Contralateral Tone and Chord. Significant simple effects of base complexity were obtained for the FUSE-1 and FUSE-BOTH Conditions ($F[2,20]=12.891$, $p<.0001$, and $F=22.044$, $p<.0001$, respectively), such that "same" responses increased with increasing base complexity. In the FUSE-1 Condition the complexity effect was consistent with an increased tendency to veridically perceive the A stimulus base ear information while fusing tone and base in the X stimulus. The FUSE-BOTH Condition effect was consistent with an increased tendency to either fuse both A and X stimuli, or veridically perceive one stimulus with migration/dominance of the other stimulus.

The minimum rate of veridical target ear perception for the A stimulus (a contralateral tone and chord) in the FUSE-1 and FUSE-BOTH Conditions was estimated (using Appendix Eq. 4) to increase as a function of increasing base complexity (top of second panel, Table 4). This also represents a minimum estimate for fusing contralateral tone and base in the X stimulus of the FUSE-1 Condition, which, as expected, is below the estimated rate of fusion for similar stimuli in the FUSE-EITHER and FUSE-NEITHER Conditions. Rather than make additional, potentially invalid assumptions to allow the estimation of a range of probabilities of potential fusion and migration/dominance of contralateral distinguishing tones in the FUSE-1 and FUSE-BOTH Conditions, conditions designed to reflect the operation of a unique perceptual organization were included in Experiment 2 to evaluate these probabilities. Clearly, however, the results of both FUSE-1 and FUSE-BOTH Conditions indicate the frequent nonveridical perception of contralateral distinguishing tones, the nature of which will be explored in greater detail in Experiment 2.

Conclusions. A few conclusions can be drawn from the results of Experiment 1. First, subjects frequently fused bases with a contralateral distinguishing tone. Due to the use of binaural chord discrimination results to correct for subject guessing, Experiment 1 is argued to provide a reliable quantification of the rate of fusion for musical DP stimuli. In addition, if TP rather than fusion had given rise to the "different" responses obtained in the FUSE-NEITHER Condition, subjects must have ignored location in making their responses (i.e., must have responded to a chord percept at an abstract position between their ears). Therefore, . . . we cannot eliminate this possibility, TP seems to be, at best, very unlikely. However, the tendency to respond based upon an inappropriate location will be evaluated in Experiment 2.

The simple effects of base complexity in the FUSE-1 Condition could be attributable to (1) an increasing tendency to appropriately perceive target ear information in the A stimulus with increasing target ear complexity, and (2) a similarly increasing tendency to fuse distinguishing tone and base of the X stimulus in the target ear. "Altered target ear perception" in the form of either migration, dominance, or fusion of contralateral distinguishing tones also was demonstrated in the FUSE-BOTH Condition to change as a function of base complexity. Although migration and dominance were not expected perceptual conditions and thus were not directly evaluated in our initial focus on fusion, altered target ear perception in the FUSE-BOTH Condition frequently might have been due to either of these perceptual tendencies, and thus warranted further investigation in Experiment 2. These instances of altered target ear perception suggest parallels with visual attention research which has investigated how people attend to and analyze stimulus features under varying degrees of stimulus complexity. These parallels will be discussed in the general discussion section.

EXPERIMENT 2: Investigating Mislocalization of Component Tones

Experiment 2 further investigated the unexpected effects of base complexity from the FUSE-1 and FUSE-BOTH Conditions of Experiment 1. No explicit empirical test existed in Experiment 1 to evaluate possible differences in altered target ear perception depending on stimulus position (A or X) within a trial. The results from the FUSE-1 Condition suggest that target ear perception of the X stimulus was more readily modified to match veridical perception of the A stimulus. Thus, the A stimulus may often serve as a perceptual template which alters schema-driven aspects of the perception of the X stimulus. Experiment 2, therefore, included conditions designed to separately assess the likelihood of a given perceptual organization (fusion, migration, or dominance) for A and X stimuli.

Experiment 1 also did not determine whether subjects fused contralateral distinguishing tones to perceive figurally bad, dissonant chords; such perception only was originally assumed on theoretical grounds to be highly unlikely. In Experiment 2, additional figurally poor, dissonant binaural and dichotic conditions were included in which both E_1 and E^b distinguishing tones were presented ipsilaterally. Dichotic conditions involving one such stimulus enabled estimation of the probability of fusion for alternative stimuli involving dichotic distinguishing tones (one of which was physically mixed with the base). Similarly, conditions were designed to determine the probability of stimulus migration or dominance by pairing stimuli containing contralateral tone and chord with a monaural major/minor chord. Finally, to better address the unlikely possibility of TP from Experiment 1,

conditions also were included to evaluate whether subjects were distracted in Experiment 1 to respond to an inappropriate location.

Method

Subjects. Musical history and performance criteria were the same as those of Experiment 1.[3] The subjects were five SUNY-Binghamton undergraduate introductory psychology students who were participating in partial fulfillment of course requirements. The first author served as a sixth subject.

Materials. In addition to the distinguishing tones, bases, and chord stimuli of Experiment 1, dissonant chords were generated by physically mixing both distinguishing tones (E and E♭) with bases of varying complexity (C-G, C-G-A, or C-G-B-D).

Procedure: Binaural Discrimination. In addition to the 80 AX binaural discrimination trials from Experiment 1, 45 randomly distributed dissonant chords trials were included to evaluate subject ability to discriminate major or minor chords from chords which included both distinguishing tones. Thirty of these trials consisted of 1 of the 6 major/minor chords from Experiment 1 (e.g., C-E-G-A, a C-major 6th chord) as the A stimulus, followed by a dissonant chord of similar complexity (C-E♭-E-G-A) as the X stimulus. Five repetitions of each AX combination were included. In the remaining 15 trials each of the 3 dissonant chords served as both A and X stimuli (5 repetitions each).

Dichotic Discrimination: Stimuli. Upon completion of the binaural discrimination trials, and after a short break, subjects performed a block of 400 randomized dichotic discrimination trials. Target ear assignment was again counterbalanced across subjects. Most stimuli were those of Experiment 1, constructed from dichotic combinations of one isolated distinguishing tone with either a base, or a chord determined to be major or minor by the inclusion of the alternative distinguishing tone. Bases and chords were of varied complexity. The remaining stimuli were monaural versions of the dissonant chords (of equally varying complexity) used in binaural discrimination; therefore, dissonant chords without simultaneous dichotic tones.

Possible perceptual organizations for each of these individual trial stimuli are again listed in Table 1. Each dichotic condition was designed with the expectation that a particular perceptual organization would result in a unique pattern of responses across conditions. In this way, dichotic conditions could be used to evaluate the tendency toward particular organizations. Only perceptual configurations of interest and corresponding responses will be discussed in the text. A complete listing of expected responses for all possible perceptual organizations and conditions (displayed for 2-tone bases and one configuration of distinguishing tones) is additionally provided in Table 2. Derivations of these responses can be obtained using Table 1 and the following description of dichotic conditions.

Dichotic Conditions. Sixty AX trials were randomly presented for each of 6 experimental dichotic conditions. Within each experimental condition, 20 trials were presented at each level of complexity. Within each level of complexity, 10 trials employed one configuration of distinguishing tones for A and X stimuli (e.g., C-E-G ‖ E♭); the other 10 trials used the opposite configuration (i.e., C-E♭-G ‖ E). Two additional control conditions of 20 trials each were included to evaluate response bias; in these trials there again was an equal probability ($p = 0.5$) for either configuration of distinguishing tones. To limit the number of dichotic trials, control conditions only involved the simplest level of target ear complexity. We now review predictions for conditions in the order of which they appear in Table 2.

The MIGRATE-1ST and MIGRATE-2ND Conditions evaluated the likelihood of subjects migrating contralateral distinguishing tones (or, alternatively, exhibited stimulus dominance effects) in the A or X stimulus, respectively. Since the X stimulus in the MIGRATE-1ST Condition and the A stimulus in the MIGRATE-2ND Condition both represent good figures (major or minor chords) with no tones presented to the contralateral ear, they were assumed to be perceived veridically. "Same" responses then would be obtained if subjects migrated contralateral distinguishing tones in the remaining stimulus. If altered target ear perception by fusion or migration of distinguishing tones occurs more frequently for X stimuli in order to match veridical target ear perception of the A stimuli, then "same" responses should occur more frequently in the MIGRATE-2ND Condition (and should increase with increasing base complexity for this condition). On the other hand, "same" responses in both conditions should similarly increase with increasing complexity if incidence of tone migration does not depend on stimulus position or order.

The DISTRACT-NO.FUSE and DISTRACT-FUSE.2ND Conditions evaluated the remote possibility that subjects in Experiment 1 were distracted, responding inappropriately to perception of tones they received other than in the target ear (e.g., responding to the contralateral ear). Distraction was evaluated by presenting bases or chords of AX stimuli to the contralateral ear rather than to the target ear. Both the DISTRACT-NO.FUSE and the DISTRACT-FUSE.2ND Conditions were generated by reversing the ears to which tones were presented in other dichotic conditions (Experiment 1's FUSE-NEITHER Condition and Experiment 2's DISSONANT-FUSE.2ND Condition described below, respectively). "Same" responses in the DISTRACT-NO.FUSE Condition presumably would result only if subjects attended to and veridically perceived the tone complex presented to the contralateral ear. "Same" responses in the DISTRACT-FUSE.2ND Condition were thought to primarily reflect responding to the contralateral ear and fusion of distinguishing tone and chord in the X stimulus. In this condition, "same" responses also could theoretically reflect (1) migration or dominance of the isolated distinguishing tone in the A stimulus, or (2) TP of the X stimulus. However, tone migration in the A stimulus seemed unlikely since it was more likely that subjects were aware that no tones were presented to the target ear. TP was also unlikely for reasons already discussed in Experiment 1.

The final two experimental conditions, the DISSONANT-FUSE.1ST and DISSONANT-FUSE.2ND Conditions, evaluated the likelihood of subjects fusing contralateral distinguishing tones to perceive dissonant chords. The evaluation was conducted individually for A and X stimuli by reversing the order of stimuli across the two conditions. Assuming veridical perception of the other trial stimulus, "same" responses would be obtained if subjects fused the contralateral distinguishing tone and chord (or, less likely, exhibited TP) of (1) the A stimulus in the DISSONANT-FUSE.1ST Condition, and (2) the X stimulus

in the DISSONANT-FUSE.2ND Condition. Based upon the previously cited Gestalt literature on figural goodness, "same" responses were not expected to occur frequently. As previously noted, increases in "same" responses as a function of increasing complexity would be consistent with decreasing figural goodness as base complexity increases; the reverse pattern should be observed if figural goodness increases with increasing complexity.

In the above experimental conditions, almost all (likely) perceptual organizations would result in "different" responses by subjects, and those organizations which would be reflected by "same" responses were not hypothesized to occur frequently. The final two dichotic conditions, therefore, were designed to consistently result in "same" responses, thus lowering the likelihood of consistent "different" responding, and acting as control conditions to determine if subjects were exhibiting a response bias. The DISTRACT-CONTROL Condition was generated from the FUSE-EITHER Condition stimuli of Experiment 1, but presented bases to the contralateral ear and distinguishing tones to the target ear. Subjects presumably would not respond "different", since such responses would reflect not only distraction to the contralateral ear, but also fusion of either the A or the X stimulus (not both).

In the NO.FUSE-CONTROL Condition, the same chord was presented to the target ear for both A and X stimuli, along with a dissonant contralateral distinguishing tone in the A stimulus. "Different" responses only would be obtained if subjects (1) were distracted to the contralateral ear, or (2) exhibited fusion or TP of the A stimulus, neither of which was hypothesized to occur.

Results and Discussion

Binaural Discrimination. Mean subject accuracy (in percent correct) is shown for each binaural condition and level of complexity in the upper panel of Table 5. Subjects again discriminated isolated tones with perfect accuracy and maintained high accuracy levels of chord discrimination.

Chord discrimination results initially were analyzed in a 4 X 3 ANOVA, with condition (same and different trials for both major/minor and dissonant chord discrimination) and complexity (number of tones) as the respective factors. The main effect of condition was not significant ($F[3,15]=.514$, $p=.6790$). There was a significant main effect of complexity ($F[2,10]=7.398$, $p=.0107$), as well as a marginal condition X complexity interaction ($F[6,30]=2.141$, $p=.0776$).

The table of means reveals that both the complexity main effect and the marginally significant interaction can be attributed primarily to decreasing accuracy with increasing complexity on different chord trials. This trend was revealed by an analysis of simple main effects of complexity for each condition, which were significant, or approached significance, for different chord trials ($F[2,10]=3.879$, $p=.057$ for major/minor chord discrimination; $F[2,10]=5.309$, $p=.027$ for dissonant chord discrimination), but failed to approach significance for same chord trials ($F[2,10]=1.404$, $p=.290$ for major/minor chord discrimination; $F[2,10]=1.400$, $p=.291$ for dissonant chord discrimination).

These results are consistent with the binaural findings of Experiment 1 in indicating an increased difficulty in perceptually isolating individual tones as the number of tones present in a complex stimulus increases. Binaural chord discrimination performance again was used to provide error rates for dichotic discrimination performance, enabling what, in theory, should be an accurate evaluation of dichotic perceptual organization.

Dichotic Discrimination. Mean percentage of "same" responses for each dichotic condition and level of complexity is shown in the lower panel of Table 5. A 6 X 3 ANOVA with experimental dichotic condition and complexity as respective factors was first conducted. All effects were significant; the main effect of complexity ($F[2,10]=13.364$, $p=.0015$), the main effect of condition ($F[5,25]=46.706$, $p<.0001$), and their interaction ($F[10,50]=6.671$, $p<.0001$).

Insert Table 5

The main effect of base complexity and the interaction can be attributed to the conditions evaluating migration, dominance or fusion of contralateral distinguishing tones, reflected by significant simple main effects of complexity for the MIGRATE-1ST ($F[2,10]=13.527$, $p<.001$), MIGRATE-2ND ($F[2,10]=11.740$, $p>.002$), and DISSONANT-FUSE.1ST Conditions ($F[2,10]=4.137$, $p=.049$). The simple main effect of complexity failed to reach significance for the DISSONANT-FUSE.2ND Condition ($F[2,10]=2.41$ $=.139$). The simple effects reveal significant overall increases in percent "same" responses with increasing complexity (a decreases in "same" responses with increasing complexity did not reach significance by Tukey-tests).

As expected, subjects almost exclusively responded "same" on DISTRACT-CONTROL trials, where all target ear components for A and timuli were identical. Therefore, subjects seemingly were not responding "different" on the basis of a simple response bias, nc were they responding to percepts at inappropriate locations, in the experimental dichotic conditions. The unexpectedly frequent "different" responses on NO.FUSE-CONTROL trials are attributed to a tendency to fuse contralateral distinguishing tones (see below).

As expected, the simple main effects of base complexity did not approach significance in the DISTRACT-NO.FUSE and DISTRACT-FUSE.2ND Conditions ($F[2,10]=.301$, $p=.747$, and $F[2,10]=1.189$, $p=.344$, respectively). Therefore, response (distraction) to the contralateral ear seldom occurred and was not made more likely by increasing complexity. Incidence of distraction across all levels of complexity, furthermore, always was within the mean error rates for subjects on binaural discrimination trials, indicating that subjects were not distracted to respond on the basis of contralateral or mislocalized information. These results are important because the incidence of TP should not exceed the rate of responding to mislocalized information.

Tukey comparisons for the MIGRATE-1ST and MIGRATE-2ND Conditions reveal that the simple effects were attributable to significant increases in "same" responses as complexity increased from 2-tone to 3-tone bases. The probabilities of migration dominance for A and X stimuli with increasing complexity, shown in panel 3 of Table 4, were calculated (using

Appendix Eq. 5a') to increase substantially from 2- to 3-tone complexity, but then to decrease slightly at the highest level of complexity. This same pattern of results was observed for both A and X stimuli. Thus, regardless of when stimuli are presented on a trial, increased migration or dominance with increasing complexity may asymptote when the number of tones is large enough to make it difficult to perceptually isolate individual tones.[6]

Because they would indicate either TP or fusion of contralateral distinguishing tones to perceive figurally bad chords, "same" responses in the DISSONANT-FUSE.1ST and DISSONANT-FUSE.2ND Conditions (as well as in the NO.FUSE-CONTROL Condition) were not hypothesized to frequently occur. The high percentage of "same" responses obtained across levels of complexity was unexpected, as was the similar tendency to respond "different" on roughly half of all trials in the NO.FUSE-CONTROL Condition. The accompanying increased tendency to respond "same" with increasing complexity on DISSONANT-FUSE trials (which was significant for the DISSONANT-FUSE.1ST Condition) further suggests that fusion increased with increasing complexity. "Same" responses in the DISSONANT-FUSE.1ST and DISSONANT-FUSE.2ND Conditions again could not have been due to TP unless subjects responded to the location of the triplex percepts rather than veridical target ear perception. However, we already have noted that subjects exhibited extremely low probabilities for responding to inappropriately localized percepts. Subjects therefore were most likely responding to target ear rather than triplex percepts. If we assume that TP is an unlikely alternative to musical DP, we can calculate the probability of fusing contralateral distinguishing tones (using Eq. 5b') to perceive a dissonant chord. As shown in panel 4 of Table 4, these probabilities, averaged across A and X stimuli, increased from 2- to 3-tone levels of base complexity and remained stable at the 4-tone level of base complexity (0.65, 0.80, and 0.79, respectively).

Means from both MIGRATE and DISSONANT Conditions from Experiment 2 were submitted to a probability formula (Appendix Eq. 6) to calculate the overall likelihood of altered target ear perception. Probabilities were calculated individually for A and X stimuli at each level of complexity and are displayed in the bottom panel of Table 4. Stimulus position (A or X) did not seem to play a critical role in target ear perception of the current stimuli, since no overall systematic bias for altered target ear perception of the X stimulus was found.

The obtained probabilities were averaged across stimulus position to obtain overall probabilities of altered target ear perception for each level of complexity. These probabilities were 0.68 for 2-, 0.93 for 3-, and 0.90 for 4-tone bases. As complexity increases, so does the likelihood of altered target ear perception. The probabilities at the two higher levels of complexity are comparable to the accuracy rates for binaural discrimination trials and are therefore essentially at ceiling.

General Discussion

Having quantified the probabilities of various perceptual organizations (i.e., fusion, migration, dominance, TP, distraction), we must provide an explanation of not only why specific organizations occurred, but also why their probabilities changed as a function of stimulus complexity. The following discussion develops general perceptual explanations for the increased tendency to alter target ear perception with increasing complexity. These include (1) an application of Gestalt principles, (2) an evaluation of the relative salience of contralateral tones based on related speech research, and (3) a comparison with a well-known attentional model which addresses similar findings with visual stimuli.

The observed probabilities of altered target ear perception cannot be due merely to subjects comparing the overall perceived similarity of stimulus configurations. If responses were based solely on the perceived similarity of A and X stimuli, then sufficient increases in complexity should have resulted in an increasing tendency to perceive different stimuli as similar. Similarity then should have affected perception of binaural and dichotic stimuli in an equivalent manner; increases in "same" responses for dichotic conditions would have been comparable to or less than the measured decrease in binaural discrimination performance for similar stimuli. This, however, was clearly not the case. Altered target ear perception in dichotic trials approached a maximal rate, whereas error rates on binaural different-chord trials remained quite low.

Alternatively, one could argue that the effects of complexity on altered target ear perception are attributable to memory decay. The current discrimination task probably does involve a strong memory component. Subjects must compare the X stimulus with a memorial representation of the A stimulus. This representation could be either (1) a trace (or image) of A, or (2) an encoded version of A, respectively equivalent to the product of "trace" and "context coding" processes (Braida and Durlach, 1986, as summarized by Macmillan, Braida, and Goldberg, 1987). Short ISIs and relatively simple stimuli (e.g., with few components) are required for maintaining an adequate memory trace; otherwise, the trace will substantially decay. Since piano tones rapidly achieve a steady-state spectral composition which remains relatively stable (apart from gradual amplitude decay) until offset, the current task would minimally require subjects to compare spectral properties at X stimulus onset with analogous properties at A stimulus offset. The minimum memory interval therefore becomes the 1.5 s ISI. Since this interval is comparable to estimates of the upper bound of echoic memory (e.g., Darwin, Turvey, and Crowder, 1972; Treisman and Rostron, 1972), the A stimulus trace could have sufficiently decayed to hamper its comparison with the X stimulus. Trace decay of veridical perception thus could appear as altered target ear perception. Furthermore, due to an increase in the number of stimulus components, trace decay should be more pronounced as base complexity increases.

Since its effectiveness is not influenced by ISI, context coding would be the more viable memory strategy. However, context coding also should be more difficult as stimulus complexity increases. Thus, effects of increasing stimulus (base) complexity are predicted not only by a greater likelihood of trace decay, but also by inappropriate context coding.

Implicitly, this memory model predicts that context coding should be easier for figurally good, and thus more easily encoded stimuli. Indeed, figurally good stimuli are generally argued to require less information to be represented, while also being more resistant to altered perception. Trace decay, reflecting one mechanism of altered perception, therefore should be less likely to affect the perception of figurally good stimuli. Thus, memory-decay may provide merely a variant of explanations

based upon changes in figural goodness as a function of complexity (as we originally hypothesized).

Gestalt Principles

Let us momentarily assume that figural goodness for chords is positively correlated with the degree of harmonic consonance produced by simultaneously presented tones. Dowling and Harwood (1986) note that chord stability should decrease with increasing dissonance between tones. Since the frequencies of adjacent tones in a musical scale do not form a simple integer ratio, adjacent tones should evoke dissonance. Dissonant chords are unstable in tonal music, requiring immediate resolution to simpler frequency ratios. Dissonance was present in both 3- and (possibly to a greater extent in) 4-tone bases. Overall dissonance increases, and therefore assumed figural goodness decreases, with increasing complexity for the current stimuli.

Three perceptual trends would be expected in the current experiments if figural goodness decreases as a function of increasing complexity. First, if bases/chords with many component tones represent weak figures, then these figures should be poorly represented by the listener. Therefore, as stimuli become more complex, and thus are represented less adequately, stimuli should increasingly be perceived as similar and also (as noted) more subject to memory limitations. Accuracy in binaural chord discrimination then should increase with increasing complexity for same chord trials and decrease for different chord trials, reflecting increased overall perceived similarity of chords as more component tones are presented. With the sole exception of same chord binaural discrimination trials in Experiment 2, this trend was observed.

Second, fusion of a contralateral tone with a base should decrease (favoring veridical perception) as complexity increases. This trend was observed in the calculated probabilities of fusion (from Eq. 3') for FUSE-EITHER and FUSE-NEITHER Conditions in Experiment 1.

Finally, when contralateral distinguishing tones are presented simultaneously (one of them physically mixed with the base), there should be an increasing tendency to alter target ear perception with increasing complexity even though the target ear received a full (but now less stable) chord. Again, this trend was confirmed by the calculated probabilities of altered target ear perception (from Eq. 7) for the MIGRATE and DISSONANT-FUSE Conditions in Experiment 2. Findings from both experiments therefore suggest that figural goodness may decrease as the number of tones (representing unique notes in a musical scale) increases. However, while invoking changes in figural goodness may allow a post-hoc explanation of complexity effects, it does not explain how target ear perception is systematically altered. We therefore consider several general theoretical alternatives.

Stimulus Dominance

Altered target ear perception might be argued to reflect stimulus dominance effects rather than the fusion/migration of distinguishing tones, since target ear perception would be the same under both conditions. Repp (1978a, b, and c) identified several dichotic pairs of CV syllables from a /ba/-/da/-/ga/ continuum which perfectly fused. Identification of fused pairs reflected the perceptual dominance of one syllable over its contralateral CV, i.e., only the dominant consonant was perceived. Repp suggested that such dominance may result from differences in relative amplitudes and frequencies of dichotic components, with subjects showing some bias (for bleats and CV syllables) to respond on the basis of the lower-frequency member of a dichotic pair.

Repp argued that stimulus dominance represents a lower-level, general auditory phenomenon. Dominance therefore should be expected for both speech and nonspeech stimuli, as Repp found for related stimuli which differed in their adherence to sounding like speech [in decreasing order of "speech-likeness", two-formant CV syllables, bleats (F2), transitions (from F1 and F2), chirps (F2 transitions), and timbres (F2 steady states)]. In the current study, dominance could have conceivably occurred when two distinguishing tones were presented to separate ears. All stimuli (tones, bases, and chords) were presented at equal amplitude. As a result, a distinguishing tone presented in isolation (to the contralateral ear) was more intense than the same distinguishing tone mixed with a base (in the target ear). Thus, isolated distinguishing tones may have been perceived as the more salient, or even the only, distinguishing tones. Furthermore, with increasing complexity the distinguishing tone mixed with a base becomes less intense relative to the contralateral distinguishing tone, but still is equal in intensity to the other tones in the chord. Thus, dominance based on relative intensity also could predict base complexity effects.

Despite this possible confound, some data from Experiment 2 are inconsistent with stimulus dominance predictions. Due to the consistent differences in component amplitudes and frequencies across conditions at each level of complexity, stimulus dominance, if observed, should have equally affected all stimuli in which contralateral distinguishing tones were presented. Monaural stimuli in the DISSONANT-FUSE Conditions (where both distinguishing tones were physically mixed with the base) should have been perceived veridically since subjects performed accurately on binaural trials involving equivalent stimuli. The remaining stimulus with contralateral distinguishing tones should have resulted in the target ear perception of a major or minor chord. Resulting "different" responses therefore should have increased with increasing complexity, directly opposite to the obtained pattern of results. The high percentage of "same" responses, therefore, could not reflect stimulus dominance. Thus, it is unlikely that the observed base complexity effects were the result of stimulus dominance.

Feature Integration Theory

An alternative explanation for the observed base complexity effects could include attentional constructs which predict systematic mislocalization and perceptual integration of stimulus components, and thus fusion of contralateral distinguishing tones. Such a model of attention, although not yet developed in the auditory literature, has been developed for visual stimuli by Treisman and her colleagues (Treisman, 1982; Treisman and Gelade, 1980; Treisman and Schmidt, 1982; Treisman and Gormican, 1988). Figure 1 outlines this model, called Feature Integration Theory (FIT). Individual features (values along a dimension, like color, shape, or size) of objects first are preattentively processed in parallel. Focusing attention at a particular location then integrates otherwise "free floating" features in order to perceive singular objects. FIT also predicts that features can be combined based upon top-down expectations of highly recurrent patterns, which probably are related to notions of figural

goodness.

Insert Figure 1

One source of critical evidence for FIT assertions is the nature of perceptual errors and the circumstances in which they arise. Attention becomes overloaded when the number of presented items, and thus the number of features, is large. If searching an array of many items for an object based on a conjunction of features, then "illusory conjunctions", incorrect combinations of existing features, are likely to result. These errors are argued to represent either random couplings of features or violations of expectations and have been shown to occur far more frequently than errors in which a feature is conjoined with one not present in the array (Treisman and Schmidt, 1982). Furthermore, subjects are not aware of these frequent illusory conjunctions.

In order to apply FIT to the current finding of tone migration, some loose assumptions must be made about what qualify as musical features. Let us assume that our auditory analyzers can preattentively identify all relevant information from the "musical array" in parallel, i.e., pitches, timbres, durations, intensities, and locations of individual tones, as well as harmonic relational information between tones. According to FIT, at lower levels of complexity (i.e., when the number of presented items is small), subjects should be relatively successful at integrating features to veridically perceive the appropriate chord in the target ear. However, if complexity is high (i.e., given many components), attention may become overloaded. Subjects then would group components along shared features in an attempt to approximate the presented stimulus configuration. One shared feature is that both E and E♭ distinguishing tones share a strong harmonic relationship with the tones in the base, combining with the base to respectively produce a major and minor chord. If grouped along this shared feature, mislocalization of E and E♭ tones would then be likely, resulting in illusory conjunctions in the form of migrations.

Fusion of contralateral distinguishing tones could be considered another form of illusory conjunction.[7] If the dissonance created by the simultaneous presentation of adjacent chroma (E and E♭) is coded as a feature, then such coding may result in perceptual fusion. Since base tones at higher levels of complexity also share dissonance and increase attention load, the likelihood of fusion would be expected to increase with increasing complexity.

Illusory conjunctions represent objects which have been resynthesized from feature labels. This seems initially contradictory to the Gestalt principles, which stress holistic perception. However, Treisman (see above) has suggested that with sufficient experience, particular combinations of primitive features (e.g., simple, highly recurring figures) are processed as emergent features. It is reasonable to assume that figurally good (e.g., major/minor) chords could act as prototypic features. Without this assumption, it is possible that the A stimulus would always act as a basis for search for the X stimulus. Subjects then would have more frequently altered target ear perception in the X stimulus than in the A stimulus. However, in the absence of any such observed tendencies, this alternative seems unlikely. Thus, figurally good chords may represent emergent features. If so, chords would be less likely to act as features at high levels of complexity, based on the instability of chords containing dissonances. Altered target ear perception yet again would seem more likely with increasing complexity.

While initial FIT applications to effects of musical complexity are far from straightforward, FIT can account for the migration and fusion of contralateral distinguishing tones, the former of which is operationally identical to visual demonstrations of illusory conjunctions. In a future submission, we intend to present results which more directly verify the applicability of FIT to audition using tasks similar to those typically used in vision. Continued investigations within the framework of FIT should provide a better specification of musical features. Invoking FIT also may allow many auditory findings to be analyzed from a new attentional perspective.

Applications and Summary of Findings

The current study suggests that research using DP stimuli can evaluate much more about auditory processing than simply addressing claims for or against modularity. Fusion incidence can be used to identify variables (e.g., complexity) which are critical to perceptual organization, as well as to reveal necessary conditions for various illusory percepts to occur.

The results represent a quantification of the probabilities of various perceptual organizations for musical stimuli as a function of stimulus complexity. Data from several conditions verified that fusion occurs at a substantial rate in musical DP stimuli, and was inconsistent with the postulation of TP by phonetic modularity supporters. Furthermore, the incidence of nonveridical perception seems to be determined in part by the figural goodness of both the fused and base percepts, and may reflect organizational tendencies which were originally demonstrated in vision. Migration and fusion of chord distinguishing tones were demonstrated to increase as a function of increasing complexity. Taken in conjunction with the slight decrease in fusion of contralateral base and tone (to perceive a major/minor chord) as a function of complexity, migration/fusion of distinguishing tones is consistent with the notion that increasing complexity decreases figural goodness. Decreasing figural goodness, therefore, is argued to decrease the likelihood of veridical perception. Furthermore, migration and fusion are consistent with feature integration as conjectured by FIT.

The application of FIT to the current results suggests a generalized attentional basis for future (DP) studies of auditory perceptual organization. Attentional factors also may affect perceptual organization in the presence of variation along other stimulus variables (e.g., component duration, intensity, or relative frequency position), which merit further research.

References

Brown, J. M. (1942). An experimental study of the phenomenon of closure as a threshold function. Journal of Experimental Psychology, 30, 273-294.

Figure 2b

Bregman, A. S. (1987). The meaning of duplex perception: Sounds and transparent objects. In M. E. H. Schouten (Ed.), The Psychophysics of Speech Perception. Boston: Martinus Nijhoff NATO-ASI Series.

Bregman, A. S. (1990). Auditory scene analysis: The perceptual organization of sound. Cambridge, MA: MIT Press.

Ciocca, V. & Bregman, A. S. (1989). The effects of auditory streaming on duplex perception. Perception & Psychophysics, 46(1), 39-48.

Collins, S. C. (1985). Duplex perception with musical stimuli: A further investigation. Perception & Psychophysics, 38, 172-177.

Cutting, J. E. (1976). Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening. Psychological Review, 83, 114-140.

Darwin, C. J., Turvey, M. T., & Crowder, R. G. (1972). An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. Cognitive Psychology, 3, 255-267.

Deutsch, D. (1974). An auditory illusion. Nature, 251, 307-309.

Dowling, W. J. & Harwood, D. L. (1986). Music Cognition. New York: Academic Press.

Fowler, C. A. & Rosenblum, L. D. (1990). Duplex perception: a comparison of monosyllables and slamming doors. Journal of Experimental Psychology: Human Perception and Performance, 16(4), 742-754.

Hall, M. D. & Pastore, R. E. (1992). Musical duplex perception: Perception of figurally good chords with subliminal distinguishing tones. Journal of Experimental Psychology: Human Perception and Performance, 18(3), 752-762.

Liberman, A. M., Isenberg, D. & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. Perception & Psychophysics, 30, 133-143.

Liberman, A. M. & Mattingly, I. G. (1989a). A specialization for speech perception. Science, 243, 489-494.

Liberman, A. M. & Mattingly, I. G. (1989b). Motor theory of speech perception revisited. Cognition, 21, 1-36.

Macmillan, N. A., Braida, L. D., and Goldberg, R. F. (1987). Central and peripheral processes in the perception of speech and nonspeech sounds. In M. E. H. Schouten (Ed.), The Psychophysics of Speech Perception (pp.28-45). Boston: Martinus Nijhoff NATO-ASI Series.

Mann, V. A., Madden, J., Russell, J. M., & Liberman, A. M. (1981). Further investigation into the influence of preceding liquids on stop consonant perception. Proceedings of the 101st meeting of the Acoustical Society of America, 69(S1), S91(A).

Mattingly, N. A. & Liberman, A. M. (1988). Speech and other auditory modules. Haskins Laboratories Status Report on Speech Research #SR-93/94, 67-84.

Nusbaum, H., Schwab, E., & Sawusch, J. (1983). The role of "chirp" identification in duplex perception. Perception & Psychophysics, 33(4), 323-332.

Pastore, R. E., Schmuckler, M. A., Rosenblum, L., & Szczesiul, R. (1983). Duplex perception with musical stimuli. Perception & Psychophysics, 33, 469-474.

Rand, T. C. (1974). Dichotic release from masking for speech. Journal of the Acoustical Society of America, 55, 678-680.

Repp, B. H. (1978a). Stimulus dominance in fused dichotic syllables. Haskins Laboratories Status Report on Speech Research #SR-55/56, 133-148.

Repp, B. H. (1978b). Categorical perception of fused dichotic syllables. Haskins Laboratories Status Report on Speech Research #SR-55/56, 149-161.

Repp, B. H. (1978c). Stimulus dominance and ear dominance in fused dichotic speech and nonspeech stimuli. Haskins Laboratories Status Report on Speech Research #SR-55/56, 163-179.

Repp, B. H., Milburn, C., & Ashkenas, J. (1983). Duplex perception: Confirmation of fusion. Perception & Psychophysics, 33, 333-357.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. Journal of Experimental Psychology: Human Perception and Performance, 8, 194-214.

Treisman, A. (1990). Variations on the theme of feature integration: Reply to Navon (1990). Psychological Review, 97(3), 460-463.

Treisman, A. & Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97-136.

Treisman, A. & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. Cognitive Psychology, 14, 107-141.

Treisman, A. & Gormican, S. (1988). Feature analysis in early evidence from search asymmetries. Psychological Review, 95(1), 15-48.

Treisman, M. & Rostron, A. B. (1972). Brief auditory storage: A modification of Sperling's paradigm applied to audition. Acta Psychologica, 36, 161-170.

Wertheimer, M. (1958). Principles of perceptual organization. In D. C. Beardslee & M. Wertheimer (Eds.), Readings in Perception. Princeton: Van Nostrand Company, Inc.

Whalen, D. A. Liberman, A. M. (1987). Speech perception takes precedence over nonspeech perception. Science, 237, 169-171.

Woodworth, R. S. & Schlosberg, H. (1954). Experimental psychology. New York: Holt.

Appendix: Probabilities for Altered Target Ear Perception

Probabilities for altered target ear perception in both experiments were calculated from several formulae. Formulae were simplified by a few basic assumptions which are specified for the given conditions. This appendix provides the derivation of formulae used for each experiment and the probabilities estimated by inserting the actual results into these formulae. Symbols used in formulae are summarized in Table 6. Each symbol reflects the probability of a given perception or response. The probability formulae, shown in Table 7, are discussed below.

Each probability was modified by corresponding accuracy/error rates on binaural chord trials, which represented a correction procedure for guessing (Woodworth & Schlosberg, 1954). In the formulae, a and b are the accuracy rates under physically same and different trials. These rates are set equal to the binaural accuracy rates for the given level of base complexity. 1-a and 1-b are the corresponding binaural error rates under physically same and different trials. For each formula, 3 calculations can be performed, one for each level of base complexity.

Insert Table 6

Experiment 1

"Same" responses in the FUSE-EITHER condition reflected correct responses, a, to fusion of neither trial stimulus, $_eF_e$, or to fusion of both stimuli, F_e, as well as an error responses, 1-b, to fusion of only one stimulus, $1-(_eF_e + F_e)$. This probability is expressed in Eq. 1, with terms collected to obtain Eq. 1b. Eq. 1 has two unknown variables, the probability of fusing both stimuli (F_e) and the probability of fusing neither stimulus ($_eF_e$). "Same" responses to the FUSE-NEITHER Condition reflect a correct response to fusing neither stimulus or an incorrect response to all other perceptual events, giving rise to Eq. 2, which has one unknown variable, $_eF_e$. The right term in Eq. 2 is equal to the bracketed term in Eq. 1b. Since stimuli in the FUSE-EITHER and FUSE-NEITHER Conditions are similar in all respects except for the isolated distinguishing tone in the X stimulus, it is assumed that the probability of fusing neither trial stimulus occurs at an equal rate for both conditions. Eq. 3 is obtained by substituting the directly measured probability, (s | FUSE-NEITHER), for the bracketed term in Eq. 2b. This formula, rewritten as Eq. 3', solves for F_e, the probability of fusing both trial stimuli in the two conditions to perceive complete chords in the target ear. This probability, F_e, is .57, .42, and .44 respectively for two-, three-, and four-tone bases. If we assume that fusion of the A or X stimulus in either condition are independent and equally probable events, the probability of fusing a single stimulus should be equivalent to the square root of the probability of fusing both stimuli, or .75, .65, and .66 as a function of increasing base complexity.

Insert Table 7

The probability of subjects responding "same" to target ear components for the FUSE-1 Condition, expressed in Eq. 4, depends on a correct response to the combination of veridical target ear perception for the A stimulus and fusing the X stimulus to perceive a complete chord in the target ear (i.e., $F_e \cdot V_e$), plus an erroneous response to all other conditions. By solving for $F_e \cdot V_e$ as a single term, Eq. 4 permitted solution for the minimum estimated probabilities of veridical A stimulus perception, V_e, and fusing the X stimulus, F_e, in the FUSE-1 Condition. This minimum incidence of V_e and F_e was estimated to be .21, .28, and .55 as a function of increasing complexity. Because the physical stimuli used as A and X in this condition are not equivalent, probabilities of any perceptual organization are not transferable across stimuli.

Experiment 2

Probabilities of migrating and fusing contralateral distinguishing tones were estimated respectively by Eqs. 5a and 5b. These probabilities were individually calculated for A (shown) and X stimuli. Formulae for X stimuli can be obtained by substituting MIGRATE-2ND and DISSONANT-FUSE 2ND for MIGRATE-1ST and DISSONANT-FUSE 1ST Conditions, respectively, as well as "," for "." symbols and vice versa.

The stimulus in these conditions (X for the displayed equations) that was presented monaurally was assumed to always be perceived veridically [above $V_e = 1$]. Substituting 1 for V_e in Eqs. 5a and 5b results in collapsed Eqs. 5a' and 5b'. As a result, the calculated probabilities for migrating contralateral distinguishing tones (using Eq. 5a') as a function of increasing complexity were .03, .62, and .59 for the A stimulus, and .16, .68, and .48 for the X stimulus. Similarly, the probabilities for fusing contralateral distinguishing tones (using Eq. 5b') as a function of increasing complexity were .65, .67, and .88 for the A stimulus, and .65, .93, and .70 for the X stimulus.

Finally, the probability of either fusing or migrating distinguishing tones, i.e., the likelihood of altering target ear perception, was individually calculated using Eq. 6 for each given stimulus (A or X, shown for A). Calculations for X stimuli can be obtained by substituting "," for "." in the equation. By substituting the probabilities of fusing and migrating contralateral distinguishing tones obtained from Eqs. 5a' and 5b' into Eq. 6, the following probabilities of altered base ear perception thus were generated respectively for two-, three-, and four-tone bases: *for the A stimulus, .66, .88, and .95, for the X stimulus, .71, .95, and .84, averaged across A and X stimuli, .68, .93, and .90.*

Acknowledgments

Based upon work supported by National Science Foundation Grant BNS8911456 and Grants F496209310033 and F49609310327 from the Air Force Office of Scientific Research. Opinions, findings, conclusions, and recommendations are the authors' and do not necessarily reflect views of the granting agencies. We gratefully acknowledge the following colleagues for their comments and suggestions on drafts of the manuscript: Richard Fahey, Xiao-Feng Li, Dawn G. Blasko, Wenyi Huang, and Jennifer L. Cho. Requests for reprints should be sent to either Michael D. Hall or Richard E. Pastore at the Department of Psychology, State University of New York at Binghamton, Binghamton, New York, 13902-6000.

Endnotes

1. Collins (1985) found a reduced rate of musical DP when contralateral components were asynchronous by 125 ms, whereas smaller asynchronies between chirp and base in DPS result in segregation (Cutting, 1976). The Collins findings could be argued to reflect the reliance on different integration cues in musical and speech DP. In speech DP, distinguishing transitions terminate with the onset of the corresponding formant in the base. Fusion is primarily based upon the good continuation of frequency as a function of time existing between transition and its corresponding formant. DPS, therefore, should be disrupted with component stimulus asynchrony. In musical DP, distinguishing tones are normally presented for the duration of the base. Fusion is based upon the relationship between component frequencies, and DP, therefore, should be less affected by small degrees of component asynchrony. It alternatively could be argued that these findings reflect differences in the use of information by speech and nonspeech systems.

2. The issue of the predicted role of figural goodness is actually more complicated than has been expressed. Clearly the fusion rate depends upon the difference in degree of closure for the base and fused stimuli. Thus, if increasing complexity decreases articulation of both base and fused stimuli, but more so for the base, then it is still possible for the rate of fusion to increase. Our current interest is in the determination of the nature of the relationship between stimulus complexity and rate of altered perception. The issue of figural goodness then will be indirectly addressed on the basis of the obtained pattern of results.

3. Data from an additional 21 subjects were discarded because they did not meet the a priori performance criterion. The large dropout rate is generally attributed to the extremely limited musical experience of the original subjects who were self-selected and often failed to read the minimal musical experience criteria for participation, although clearly there also was an occasional inattentive subject.

4. It is now generally accepted that the syllable percept in DPS is the result of a perceptual fusion of chirp and base (Repp, Milburn, and Ashkenas, 1983) rather than a cognitive integration of the separately perceived and presumably categorized components (Nusbaum, Schwab, and Sawusch, 1983). The argument in favor of perceptual fusion provides a strong basis for assuming that the probability of responses should be equal for physically mixed (binaural) and perceptually mixed (fused) stimuli.

5. There was again a high drop-out rate (eighteen subjects) due to the inability of subjects to accurately discriminate binaural major and minor chords and thus meet our a priori criteria for participation.

6. It could be argued that the consistently obtained maximum probability of altered target ear perception (via migration or fusion of contralateral distinguishing tones) for 3-tone, rather than 4-tone base stimuli, reflects minimal figural goodness for 3-tone base stimuli. In the current major/minor sixth chords, the two highest frequency components are not widely separated in frequency in relation to component tones for either 2- or 4-tone bases. As a result, overall perceived dissonance for 3-tone bases may actually exceed that for 4-tone bases, possibly making altered target ear perception more likely despite the reduced level of base complexity.

7. Although fusion in DP could be considered to be an example of an illusory conjunction, our current focus is more narrow.

Anticipated Perceptual Structures for Stimuli

Stimuli	Fusion	Migration	Dominance	Triplex Perception
t.e. ‖ c.e.	t.e. ‖ c.e.	t.e. ‖ c.e.	t.e. ‖ c.e.	t.e. | between | c.e.
C-G ‖ E	C-E-G ‖ E	-	-	C-G | [C-E-G] | E
C-G ‖ E♭	C-E♭-G ‖ E♭	-	-	C-G | [C-E♭-G] | E♭
C-E-G ‖ E♭	C-E♭-E-G ‖ E♭	C-E♭-G ‖ E	C-E♭-G ‖ E	C-E-G | [C-E♭-E-G] | E♭
C-E♭-G ‖ E	C-E♭-E-G ‖ E	C-E-G ‖ E♭	C-E-G ‖ E	C-E♭-G | [C-E♭-E-G] | E
* E ‖ C-G	E ‖ C-E-G	-	-	E | [C-E-G] | C-G
* E♭ ‖ C-G	E♭ ‖ C-E♭-G	-	-	E♭ | [C-E♭-G] | C-G
*C-E♭-E-G ‖	-	-	-	-
* ‖ C-E♭-E-G	-	-	-	-

*Additional Experiment 2 stimuli.

Table 1. Possible perceptual organizations for two-tone base stimuli presented in dichotic discrimination trials. Three-and four-tone bases can be attained by respectively adding "A" and "B-D" to the C-G base shown. Double lines represent a midpoint between the two ear locations. Target ear (t.e.) stimuli are displayed to the left of lines; contralateral ear (c.e.) stimuli are displayed to the right.

Possible Responses and Perceptual Organizations for Dichotic Conditions

Condition	Stimuli		Predicted Responses [same (s)/ different (d)]			
	A Stim.	X Stim.	Veridical/ TP in t.e.	Fusion/ triplex percept	Migration/ Dominance	Distraction to c.e.
	t.e. ‖ c.e.	t.e. ‖ c.e.		A X Both	A X Both	
Experiment 1:						
FUSE-EITHER	C-G ‖ E	C-G ‖ E	s	d d s	- - -	s
FUSE-NEITHER	C-G ‖ E	C-G ‖ E♭	s	d d d	- - -	d
FUSE-1	C-E-G ‖ E♭	C-G ‖ E	d	d s d	d - -	d
FUSE-BOTH	C-E-G ‖ E♭	C-E♭-G ‖ E	d	d d s	s s d	d
Experiment 2:						
MIGRATE-1ST	C-E♭-G ‖ E	C-E-G ‖	d	d - -	s - -	d
MIGRATE-2ND	C-E-G ‖	C-E♭-G ‖ E	d	- d -	- s -	d
DISSONANT-FUSE.1ST	C-E-G ‖ E♭	C-E♭-E-G ‖	d	s - -	d - -	d
DISSONANT-FUSE.2ND	C-E♭-E-G ‖	C-E-G ‖ E♭	d	- s -	- d -	d
DISTRACT-NO.FUSE.	E ‖ C-G	E♭ ‖ C-G	d	d d d	- - -	s
DISTRACT-FUSE.2ND	‖ C-E♭-E-G	E♭ ‖ C-E-G	d	- d -	- d -	s (w/fuse X)
DISTRACT-CONTROL	E ‖ C-G	E ‖ C-G	s	s s s	- - -	s
NO FUSE-CONTROL	C-E-G ‖ E♭	C-E-G ‖	s	d - -	d - -	d

Table 2. Dichotic conditions for both experiments and possible corresponding responses given various perceptual organizations of the stimuli. Again, stimuli are displayed only for two-tone bases and one combination of distinguishing tones.

Type of Trial	Condition	Base Complexity 2-tone	3-tone	4-tone
		Discrimination Performance (Percent Correct)		
Binaural	Same tone	100.0 (0)		
	Different tone	100.0 (0)		
	Same chord	89.8 (3.6)	92.6 (2.4)	97.0 (3.0)
	Different chord	92.2 (2.8)	90.9 (3.2)	81.8 (4.6)
		(Percent "Same" Responses)		
Dichotic	FUSE-EITHER	98.2 (1.0)	97.7 (1.8)	97.3 (2.3)
	FUSE-NEITHER	51.7 (10.0)	62.6 (7.3)	62.8 (7.4)
	FUSE-1	25.3 (7.6)	32.8 (6.1)	61.4 (6.2)
	FUSE-BOTH	19.7 (4.5)	50.0 (7.5)	66.1 (8.9)

Table 3. Discrimination performance for binaural (percent correct) and dichotic trials (percent "same" responses) for subjects in Experiment 1. Standard errors are shown in parentheses.

Table 4. Calculated probabilities of fusion and migration for Experiments 1 and 2.

Probability	*Base Complexity (# tones)* 2	3	4	*Equation*
Stimuli: Base | Distinguishing Tone (e.g., C-G | E)				
EXPERIMENT 1:				
Fuse A and X	.57	.42	.44	3'
Fuse A or X	.75	.65	.66	3'
Stimuli: Base + Tone | Other Tone (e.g., C-E-G | F)				
EXPERIMENT 1:				
Veridical Perception (minimum estimate)	.21	.28	.55	4
EXPERIMENT 2:				
Migrate A	.03	.62	.59	6a
Migrate X	.16	.68	.48	6a
Mean	.10	.65	.54	
Fuse A	.65	.67	.88	6b
Fuse X	.65	.93	.70	6b
Mean	.65	.80	.79	
Migrate or Fuse A	.66	.88	.95	7
Migrate or Fuse X	.71	.98	.84	7
Mean	.69	.93	.90	

Type of Trial	Condition	Discrimination (Percent Correct) Base Complexity 2-tone	3-tone	4-tone
Binaural	Same tone	100.0 (0)		
	Different tone	100.0 (0)		
	Same chord	91.7 (4.0)	95.0 (3.4)	86.5 (6.1)
	Different chord	94.2 (2.6)	96.7 (2.1)	83.2 (8.0)
	Same dissonant chord	93.3 (6.7)	87.8 (5.8)	93.3 (6.7)
	Different dissonant chord	93.6 (4.7)	86.7 (3.0)	75.0 (4.3)
		(Percent "Same" Responses)		
Dichotic	MIGRATE-1ST	9.0 (3.7)	59.7 (8.8)	65.0 (13.9)
	MIGRATE-2ND	20.6 (5.1)	63.8 (5.8)	57.9 (12.6)
	DISTRACT-NO.FUSE	1.8 (1.8)	4.1 (2.3)	3.7 (3.7)
	DISTRACT-FUSE.2ND	9.8 (7.0)	4.5 (2.5)	6.3 (4.1)
	DISSONANT-FUSE.1ST	62.6 (10.3)	63.2 (7.6)	85.0 (6.9)
	DISSONANT-FUSE.2ND	63.1 (7.0)	82.7 (3.3)	72.5 (8.1)
	DISTRACT-CONTROL	97.9 (1.3)		
	NO.FUSE-CONTROL	50.7 (6.0)		

Table 5. Discrimination performance for binaural (percent correct) and dichotic trials (percent "same" responses) for subjects in Experiment 2. Standard errors are shown in parentheses.

Table 6. Symbols used in probability formulae.

Key to Equation Symbols	
Probability Symbol	(Probability of) Event
F	fusion
F_A	fusion of A stimulus*
F_X	fusion of X stimulus
F_{AX}	fuse both A and X
$_{\sim}F$	*no fusion*
$_{\sim}F_A$	no fusion of A
$_{\sim}F_X$	no fusion of X
$_{\sim}F_{AX}$	fuse neither
M	migration (or dominance)
$_{\sim}M$	no migration
V	veridical perception
s | Condition	same response | Condition
d | Condition	different response | Condition
a	binaural accuracy, identical chord trials
b	binaural accuracy, different chord trials

*Subscripts for A and X stimuli also apply to migration and veridical perception.

Table 7. Probability formulae for determining the likelihood of various perceptual organizations for the given conditions in Experiments 1 and 2.

Equation	Formulae
1	$(s \mid \text{FUSE-EITHER}) = a(_{\sim}F_{AX} + F_{AX}) + (1-b)[1 - (_{\sim}F_{AX} + F_{AX})]$
1b	$(s \mid \text{FUSE-EITHER}) = [a \bullet _{\sim}F_{AX} + (1-b)(1-_{\sim}F_{AX})] + (a+b-1) \bullet F_{AX}$
2	$(s \mid \text{FUSE-NEITHER}) = a \bullet _{\sim}F_{AX} + (1-b)(1 - _{\sim}F_{AX})$
3	$(s \mid \text{FUSE-EITHER}) = (s \mid \text{FUSE-NEITHER}) + (a+b-1) \bullet F_{AX}$
3'	$F_{AX} = (s \mid \text{FUSE-EITHER} - s \mid \text{FUSE-NEITHER})/(a+b-1)$
4	$(s \mid \text{FUSE-1}) = a \bullet F_A \bullet V_X + (1-b)(1- F_A \bullet V_X)$
5a	$(s \mid \text{MIGRATE-1ST}) = a \bullet M_A \bullet V_X + (1-b)(1 - M_A \bullet V_X)$
5b	$(s \mid \text{DISSONANT-FUSE-1ST}) = a \bullet F_A \bullet V_X + (1-b)(1 - F_A \bullet V_X)$
5a'	$(s \mid \text{MIGRATE-1ST}) = a \bullet M_A + (1-b)(1 - M_A)$
5b'	$(s \mid \text{DISSONANT-FUSE-1ST}) = a \bullet F_A + (1-b)(1 - F_A)$
6	$(F_A \text{ or } M_A) = F_A + M_A - F_A \bullet M_A$

Figure Caption

Figure 1. Proposed processes leading to object identification according to Feature Integration Theory, including preattentive independent feature abstraction and attentive conjoining of features.

Feature Integration Theory

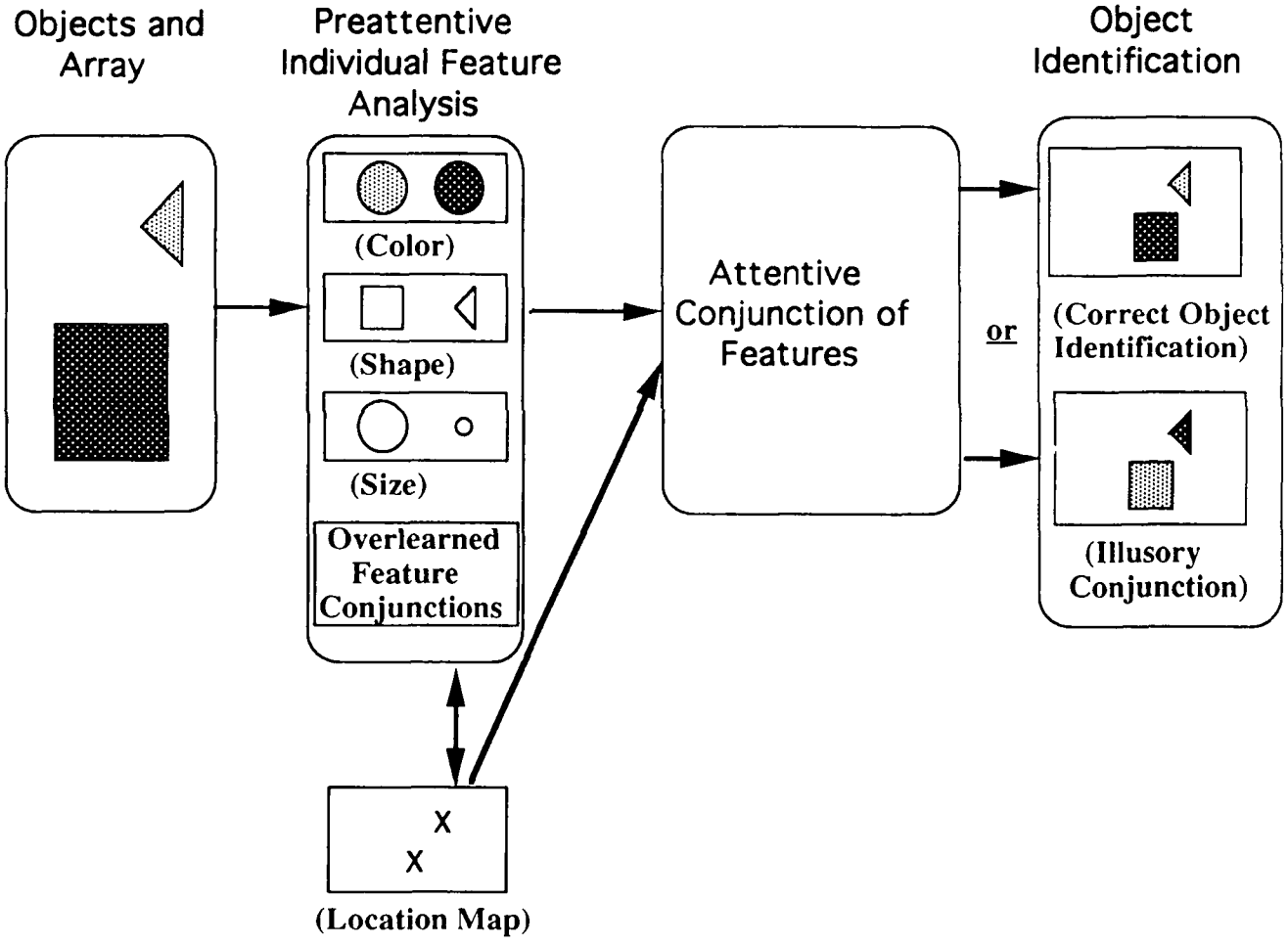


Figure 1

Mapping Percepts in the Major Variant of the Octave Illusion

Wenyi Huang, Michael D. Hall, & Richard E. Pastore
Center for Cognitive and Psycholinguistic Sciences
State University of New York at Binghamton
Binghamton, NY 13902-6000

Abstract

The current study investigated stimulus and perceptual factors critical to the commonly perceived variant of the octave illusion (e.g., Deutsch, 1974a). In contrast to the more typical illusion pattern the fused percept shifts only slightly in pitch with the corresponding shift in lateralization. The perceptual characteristics of this version of the illusion reflected properties of dichotic fusion, rather than the effects of some sequential characteristics of the stimuli. Perception of the illusion remained stable despite large variation in ISI (100-2200 ms) between dichotic pairs. Additionally, both the lateralization tendency and the pitches of fused percepts were generally not affected by change in the length of the sequence (2-, 4-, and 12-pair sequences); these tendencies also were observed with a single isolated pair of dichotic tones. Furthermore, mappings of the perceived illusory pitches consistently corresponded to other than the frequency presented to either ear, and seem to reflect a specific weighted averaging of component stimuli. The possible higher levels of processing and individual differences in the perception of this variant of the illusion also are discussed.

The octave illusion (Deutsch, 1974a) is an example of perceptual errors whose delineation may help us understand certain perceptual processes (e.g., how dichotic information is processed). The physical condition for the illusion involves the dichotic presentation of two tones separated in frequency by an octave (typically 400 and 800 Hz); the tones are rapidly alternated in each ear, such that when one ear receives the low (400 Hz) tone, the other ear simultaneously receives the high (800 Hz) tone. This physical condition is illustrated in the left column of Figure 1.

Insert Figure 1 about here

According to the original report by Deutsch (1974a), several illusory patterns were perceived. The most often perceived pattern, reported by 58% of right-handed listeners, was of "a single tone oscillating from ear to ear, whose pitch also oscillated from one octave to the other in synchrony with the localization shift" (p. 308). The perceived pitches of this pattern were verified by two subjects with absolute pitch, who identified the oscillating pitches as G_4 (392 Hz) and G_5 (784 Hz). This perceptual condition, illustrated in the right column of Figure 1, is what is called to mind when one mentions the octave illusion, and is the form of the illusion most typically investigated (see below). The second most commonly perceived illusion pattern reported by 25% of right-handed listeners was "of a single tone oscillating from ear to ear, whose pitch either remains constant or shifts very slightly" (p. 308).

Later studies by Deutsch and colleagues focused on only the most typical illusory percept, studying factors that are potentially critical to the incidence of the illusion (e.g., Deutsch 1974a, 1974b, 1975a, 1975b, 1976, 1978a, 1978b, 1981, 1988; Deutsch and Roll 1976). For example, Deutsch (1976) reported a significant tendency for right-handed listeners to hear a sequence in which pitch corresponded to the stimulus delivered to the right ear (i.e., a "right ear dominance") and lateralized to the ear receiving the higher frequency (called lateralization-by-frequency effect, also see Deutsch 1983). Based upon evidence of ear dominance and lateralization-by-frequency, Deutsch and Roll (1976) hypothesized that the illusion involves two different mechanisms. A "where" mechanism localizes the fused tone in the ear receiving the higher frequency input. A "what" mechanism determines the perceived pitch of that tone based upon the frequency presented to the dominant ear (for most right-handers, the right ear). In Deutsch (1980), subjects were selected on the basis of hearing the typical illusion pattern (as shown in Figure 1); subjects reported hearing an octave difference between successive tones. Furthermore, when asked to match the successive pitches in the illusion sequence with those of a single binaural sequence, the matches all approximated a succession of tones that were spaced an octave apart (Deutsch, 1983).

The lateralization-by-frequency effect has been studied under a number of different stimulus conditions for the group of listeners exhibiting the most common perceptual form of the illusion. Lateralization in the illusion appears to not normally depend on loudness differences between dichotic components. In fact, for some subjects, lateralization still occurred when the 800 Hz tone was substantially lower in amplitude than the 400 Hz tone. However, lateralization does depend on sequential relationships inherent in the illusion stimuli. For example, an illusion sequence of sufficient length must be presented to observe a strong lateralization-by-frequency, which is substantially weaker given 2 as opposed to 20 dichotic pairs (Deutsch 1983). Furthermore, the strength of this effect was observed to decrease as a function of increasing time between onsets of successive dichotic pairs (e.g., interstimulus interval, see Deutsch, 1982). Deutsch (1978, 1981) also demonstrated that lateralization to the higher frequency signal occurred even when the lower frequency signal was more than 12 dB greater in amplitude.

Sequential interactions in the stimulus sequence also are critical to the observance of ear dominance effects in the more typical form of the illusion (Deutsch 1980). Specifically, the size of the ear dominance is reduced when the alternating presentation of the octave component stimuli is disrupted by inserting either a binaural 599 Hz tone or a dichotic pair of different (non-octave) component frequencies between dichotic 400 and 800 Hz pairs. Finally, subjects in these studies often were selected on the basis of exhibiting a strong pitch memory. Thus, it is possible that Deutsch may have unintentionally selected subjects who perceived the illusion in a unified manner (maybe the most common manner). We will return to this issue

shortly below.

The Purpose of the Current Study

While (as noted above) factors critical to the illusion have been studied extensively for listeners who perceive a single tone with an octave difference in pitch (the most commonly reported pattern in the original study), these factors (including sequential relationships) have not been studied for the many listeners who perceive a slight (or no) pitch shift (the second most commonly perceived pattern for right-handed listeners). As an unfortunate result, explanation for the vast perceptual difference between listeners given the illusion sequence are currently not available.

We initially set-up a laboratory demonstration (described in more detail below) in which the physical conditions for the illusion were followed by the physical conditions which matched that described for the most common form of the illusion (see Figure 2). Based on the informal listening among our laboratory staff and among visitors to the laboratory, we found two basic patterns of perception which could easily be judged when using the context of the physical match to expected perception as a standard or reference. Listeners with extensive musical experience (as with Deutsch (1980) listeners who exhibited strong pitch memory) tended to hear a sequence of a pair of octave related stimuli which switched from ear-to-ear, and a faint stimulus which was more centrally localized and which seemed to shift slightly in pitch. Most listeners without musical training, instead, reported hearing a unified (probably complex) stimulus which appeared to shift somewhat toward each ear and with a small pitch shift. This latter pattern of illusion perception corresponds to the "single pitch" group originally identified by Deutsch (1974a). The current study focuses on this population and perceptual pattern. Specifically, we studied the effect of sequential relationships on both the lateralization-by-frequency effect and the perceived illusion pitches for this particular group of listeners. In mapping out a reasonable approximation of illusion perception for this alternative group of listeners, the current study provides the basis for a clear description, and potential explanations, of individual differences associated with the illusion.

It is noteworthy that the unusual and unexpected pattern of pitch results is characteristic of one type of listeners, and not of other types of listeners. As a result, we will not argue that the perceived pitches reported here primarily reflect peripheral mechanisms for frequency encoding (which are readily addressed by existing pitch models). Rather, we will argue that the difference in reported pitches across listeners in the original (Deutsch, 1974a) study probably reflect a difference in listening strategies which may vary as a function of musical experience and proficiency in processing acoustic stimuli.

Experiment 1

Deutsch reported that lateralization-by-frequency decreased with increasing time between onset of the identical frequencies at the two location (Deutsch, 1982, p. 114). Deutsch also has reported that "the durations of the tones themselves do not appear of importance and neither does the time interval between the offset of one tone and the onset of its successor" (Deutsch, 1980, p. 585). Thus, sequential characteristics affect incidence of the octave illusion.

Experiment 1 sought to replicate Deutsch's initial findings for the alternative group of subjects by systematically manipulating ISI in the octave illusion. This initial experiment used a structured self-report procedure. The subsequent experiments augmented this type of procedure with other procedures to investigate the role of various factors which may have contributed to the octave illusion.

Method

Subjects. Four undergraduate students from the State University of New York at Binghamton participated in the experiment for course credit. Another eight subjects from the university area were paid for their participation. All twelve subjects reported normal hearing. Subjects in this experiment were run in a sound-isolated booth. Each subject was evaluated in terms of the pattern they perceived given the illusion sequence.

Stimuli. 400 Hz and 800 Hz pure tones were computer generated using a 12-bit D/A converter with 10 kHz sample rate and 4 kHz low-pass filtering. The 250 ms stimuli began and ended at positive waveform zero-crossings with no ramps. All the stimuli used in subsequent experiments were generated in an identical fashion.

Procedure. Four pairs of 400 Hz and 800 Hz pure tones were presented to subjects through TDH49-10Z headphones at 75dB in the manner summarized in the left column of Figure 2, with the 800 Hz tone always presented first in the right ear. For each trial ISI was selected randomly from values ranging between 100 and 2200 ms in increments of 300 ms; ISI varied between trials. Note that the period of repetition of the stimuli (period = 2*(duration+ISI)) co-varied with ISI.

Insert Figure 2 about here

The experiment consisted of 40 trials (5 repetitions at each ISI) for each subject. On each trial subjects listened to the stimulus sequence, then indicated the nature of their perception by checking one of the six specified patterns summarized in Table 1. Subjects also were encouraged to report all perceptions which did not correspond to any of the six patterns. After the experiment, subjects also were asked to describe what they typically perceived.

Insert Table 1 about here

Results and Discussion

For all values of ISI, 10 out of the 12 subjects always reported perceiving an isolated high pitch on the right side of their head alternating with a low pitch on the left side (equivalent to the right column of Figure 2 and pattern 1 in Table 1). One of the remaining 2 subjects simply reversed pitch location, always reporting a low pitch on the right and a high pitch on the left (pattern 2) for all values of ISI. The remaining subject failed to hear the illusion, consistently reporting two simultaneous

pitches (pattern 6) to opposite ears, again for all values of ISI. In the post-experiment de-briefing, most subjects verbally indicated that the pitches were not fully lateralized to either ear.

The finding that perception of the illusion is independent of ISI (for ISI over the range of 100 to 2200 ms) indicates that the octave illusion is not sensitive to the silence gap between tones. Had ISI been critical, the illusion should have weakened or disappeared at long values of ISI.

Experiment 2

With the initial results from Experiment 1, we attempted to develop a type of objective 2IFC task tailored to investigate the location and pitch of the percepts in the octave illusion. Initially, we set up a two component trial structure, with each component consisting of a sequence of tones. One component on the trial was a sequence of dichotic 400 and 800 tones changing from ear to ear (the illusion stimulus configuration). The other consisted of a sequence of only the 800 Hz tone presented to the right ear alternated with the 400 Hz tone presented to the left ear (i.e., following the typical description of illusion perception, as summarized in Figure. 1). The effort to develop a relevant 2IFC task failed, but did succeed in demonstrating some unexpected aspects of the illusion. With an external standard as a reference on each trial, most of our laboratory staff perceived the actual location of each perceived pitch toward the ear which received the 800 Hz tone.[1] However, the perceived pitches also were quite different from (and intermediate to) the monaural 800 and 400 Hz tones. This pattern of perception is quite different from the typical conception of the illusion, but is consistent with the pattern described by Deutsch (1974a) as perception of "single pitch" group.

The subsequent experiments all build upon our informal observations that the perceived pitches and locations in the illusion definitely do not correspond to an 800 Hz tone in one ear and a 400 Hz tone in the other, but rather are relatively smaller shifts, with perceived pitch being somewhere within the octave (i.e., the high pitch was much lower than an 800 Hz tone and the low pitch was higher than 400 Hz). Obviously, the perception of a single, intermediate pitch was quite unexpected, and does not seem to have a direct counterpart in the extensive research on pitch perception. Experiment 2 was designed to quantify our informal results. Our goal was to specify what pitches are perceived when this form of the illusion is experienced. (i.e., a single tone with small alternating pitch and location). With this mapping, we can begin to address possible explanations for the unexpected findings.

Method

Subjects. Sixteen subjects participated in Experiment 2. Ten were the subjects from Experiment 1 who had perceived the expected illusion pattern. These 10 subjects participated in the first pitch-matching condition (see the procedure below). Six additional subjects were paid for participation in the subsequent conditions. All subjects reported normal hearing and, when tested (using the procedure from Exp. 1), reported perceiving the alternative form of the illusion.

Stimuli. Seventeen pure tones, ranging from 400 to 800 Hz in 25 Hz steps, were generated as stimuli. These stimuli also were used in the subsequent experiments.

Procedure. There were three separate pitch-matching conditions in this experiment. The first condition used a version of the Levitt (1971) up-down procedure to roughly match the pitch of a single binaural tone to the perceived high or low pitch. In this condition, the fixed sequence of illusion stimuli were presented as in Experiment 1, but now ISI was reduced to 50 ms and the sequence was followed by a 250 ms delay, which then was followed by a single 250 ms comparison tone. For a given block of trials, each subject was instructed to compare either only the perceived high or low illusion pitch with the pitch of the comparison tone. Subjects indicated whether the comparison pitch was higher or lower than the illusion pitch by pressing a corresponding key (up or down arrow) on a computer keyboard. The process then was repeated for the alternative illusion pitch. The order of matching high or low illusion pitch was counterbalanced across subjects.

For each block of trials, there were one up sequence and one down sequence of comparison tones which started respectively at frequencies of 400 Hz and 800 Hz. On each trial, the computer imposed the current comparison tone by random selection of one of the two sequences. If the comparison tone from that sequence was judged to be higher (or, alternatively, lower) than the illusion pitch, the frequency of the comparison tone was decreased (or increased) by 25 Hz for the next selection of the given sequence. For each sequence, the frequency of a comparison tone was recorded when the direction of frequency change of the adaptive procedure reversed sign. Each reversal in the direction of frequency change (increment or decrement) is assumed to indicate a crossing of the frequency match (point of subjective equality) with the perceived illusion pitch. Thus, in such adaptive procedures most reversals are made in the frequency region of pitch equality. A block of trials ended only after both sequences reached a minimum of 13 recorded frequencies at which the response (and frequency direction) changed. The derived statistical measures followed standard up-down procedures. In calculating the mean and standard error, the first 3 recorded frequencies were excluded, thus focusing on the measurements which were narrowly concentrated around the perceived illusion pitch. Thus, for a given pitch, each block of trials generated two mean frequency values (one for each sequence) based upon a minimum of 10 measurements.

The up-down procedure has a built-in criterion for consistency. If responses drive 2 sequences together, the subject must have a consistent basis for responding. Furthermore, step size determines precision of measurement. If all subjects give highly similar results, then subjects have similar basis for responding. Thus, results of individual subjects were examined to insure that within each block of trials the two sequences converged quickly on a specific frequency and that convergence was maintained. Across all subjects, the average difference between the two sequence means was 5.4 Hz, with an average standard deviation for sequence reversal of 10 Hz. Therefore, individual subjects were highly consistent in responding based upon a single, stable pitch percept.

The up-down procedure also has several disadvantages for our use in the current study. Because of our laboratory

structure, we could only run one subject at a time. Furthermore, because of hardware constraints, the subject had to be seated near the computer and thus in the laboratory control room (rather than in a sound chamber), shutting-down the rest of the laboratory. Therefore, for the remainder of the study we utilized a different procedure with multiple subjects being simultaneously run in sound chambers.

The second pitch-matching condition again evaluated the perceived high and low pitches in 4-pair illusion sequences, this time using the method of constant stimuli. In addition to the pragmatic concerns summarized above, there were two reasons to run this condition. Although the method of constant stimuli lacks an inherent internal standard for reliability, the shape and slope of the psychometric function of each individual can provide basis for evaluating reliability and precision of measure. Thus, the method of constant stimuli allows an alternative evaluation of the stability of perception in terms of the slope of psychometric function of each individual subject (computed by linear regression of z-score transformed data). Also, we could compare results across methods to determine if the obtained pitch matches depend upon the psychophysical method used.

The task for the subjects in all conditions was to report whether the pitch of the comparison tone was higher or lower than the designated (high or low) pitch in the illusion sequence. Subjects indicated their response by pressing one of two corresponding buttons. In this condition the comparison stimulus for each trial was randomly selected from the full octave range (400-800 Hz), not just a narrow range of comparison frequencies determined by previous trials. (Pilot work with both methods also had utilized stimuli outside of the octave, but the pitch of these stimuli was clearly highly discrepant from any illusion pitch, and their inclusion decreased the amount of relevant data which could be collected from each individual subject. The comparison stimuli thus were limited to the octave range). A third pitch-matching condition was run with method of constant stimuli to evaluate the possibility that the perceived difference between high and low pitches was a function of sequence length. This third condition used 12 pairs of illusion stimuli.

Results and Discussion

The pitch comparison results from the first condition (adaptive procedure) are summarized in Figure 3, which plots frequencies of tones matched to the high versus low pitch for each subject. The two axes of this scatter-plot represent the full range of the octave. Most of the data are concentrated in the lower half of the octave, showing that the perceived high and low pitches were more toward 400 Hz than 800 Hz. The regression line has a slope of 0.91 ($r^2 = .78$), indicating a relatively constant difference between high and low pitch matches and consistency of data across subjects. The mean frequencies of high and low pitch matches were 550 and 501 Hz, respectively, resulting in a mean difference of 49 Hz.

Insert Figure 3 about here

Under the method of constant stimuli, the psychometric functions of each individual subject in the 4- and 12-pair conditions exhibited very steep slopes: the mean frequency change for the inter-quartile range (middle 50% of a psychometric function) was only 10 Hz. The steep slope of the individual psychometric functions indicates that the perceived frequency of the high and low pitches in the octave illusion was highly stable for each subject.

The scatter-plotted results for the method of constant stimuli are summarized by the filled symbols in Figure 4 (the open symbols are the results from a subsequent experiment). The filled circles and filled squares represent the data for 4- and 12-pair sequences respectively. The pitch-match data for both sequences again are concentrated in a very narrow range of frequencies in the lower half of the octave, revealing that most subjects perceived highly similar pairs of pitches. In fact, the range of perceived pitch across subjects is too narrow to compute a meaningful regression line for the full data set. With the 4-pair sequence, respective mean frequencies of the high and low pitch match across subjects were 548 and 510 Hz when computed from the psychometric function using a linear regression based upon frequency (or 546 and 507 Hz when based upon logarithmic frequency), resulting in a mean difference of 38 Hz. With the 12-pair sequence, the mean frequencies for high and low pitches were 538 and 497 Hz (or 536 and 495 Hz when based upon logarithmic frequency), resulting in a mean difference of 41 Hz. (The lack of differences between matches obtained from regressions based on linear and logarithmic frequency reflects the steep slopes of the individual psychometric functions.) The difference between perceived high and low pitches within each condition was statistically significant in all three pitch-matching conditions [4-pair adaptive procedure, $t(9) = 5.96$, $p < .01$; 4-pair constant stimuli procedure, $t(5) = 5.24$, $p < .01$; and 12-pair constant stimuli, $t(5) = 5.20$, $p < .01$].

Insert Figure 4 about here

Given the significant overlap of the pitch-matching data for 4- and 12-pair constant stimuli conditions, it is not surprising that the t-test of the perceived pitch difference between the 4-pair and 12-pair sequence was far from being statistically significant [$t(5) = .20$, $p > .50$]. Perceived pitch, therefore, was not a function of sequence length, at least for the sequences studied. Furthermore, the difference between the perceived pitches measured using the adaptive procedure (first condition) did not differ from that measured using the method of constant stimuli (second condition) [$t(14) = .94$, $p > .2$], showing that method is not important in the precision of pitch measurement. Both procedures consistently indicate pitches correspond to narrow range between approximately 500 and 550 Hz, and definitely a less than the 400 Hz defining the full octave. Therefore, similar pattern of percepts were observed in either method. The following two experiments used a single method, the method of constant stimuli.

We do not claim that our subjects, who are relatively musically naive, are perceiving a singular tonal quality. It is quite possible that the subjects are hearing a complex stimulus which has a distinct pitch quality, and that this pitch clearly does not correspond to the pitch of either of the original octave stimuli. Instead, subjects reported perceiving pitches which were

intermediate between the original frequencies and which were skewed toward the low-frequency end of the octave. The pitch match results seem to suggest that our perceptual system is performing some type of weighted average of the dichotic frequency inputs, resulting in the perception of a fused stimulus of intermediate pitch. A quantification of this averaging process will be presented in the general discussion along with a discussion of the possible nature of the pitch percept.

Experiment 3

A reviewer on an earlier version of this manuscript suggested that our subjects were actually hearing stimuli an octave apart, with the pitch results reflecting a lack of ability of musically naive subjects to perform the matching task. With this lack of subject ability, the two procedures were conjectured to exhibit a regression toward an intermediate frequency, rather than providing separate measure of the same percept. This reviewer further suggested that the use of a broader range of comparison stimuli probably would result in different pitch matches. A second reviewer conjectures that the intermediate pitch results might be artifacts of having used linear rather than log frequency for the psychometric functions. Therefore, in this experiment we again will present both solutions. Experiment 3 represents a control to verify that the pitch results are reasonable, and not artifacts.

Method

Subjects. Ten subjects participated in Experiment 3. Nine were undergraduate students who participated in fulfillment of course requirements. The first author also participated.

Stimuli. The comparison tones were generated in the same fashion as in the earlier experiments. The frequencies of the comparison tones ranged from 350 to 850 Hz with a 25 Hz step size.

Procedure. Two conditions were run using the method of constant stimuli. In the first condition, subjects matched the high or low pitch of a 4-tone monaural sequence consisting of alternating 400 and 800 Hz tones, thus mimicking the most typical illusory pattern of pitches reported by Deutsch (1974a). The second condition instead used the 4-pair illusion sequence of Experiment 2 as stimuli.

Results and Discussion

The pitch-match results from both conditions are shown in Figure 5. The open symbols represent the pitch matches from the monaural condition; the filled symbols represent pitch matches from the illusion sequence. There is no overlap of matches between the two conditions. The dark asterisk represents the mean pitch matches obtained in Experiment 2 for the the 4-pair illusion using the method of constant stimuli. This reference point is well within the range of results for the replication of this condition.

The mean pitch match for the monaural 400 and 800 Hz tones were 444 Hz and 739 Hz, respectively; the difference of 295 Hz between high and low pitches is significant [$t(9)=13.25$, $p<.01$]. In the illusion sequence condition, the mean matches for the high and low pitches were 528 and 479 Hz respectively; the difference of 49 Hz between pitches is also significant [$t(9)=3.68$, $p<.01$]. Clearly, the actual pitch matches and the magnitude of differences between high and low pitches differ significantly across the two conditions [295 vs. 49 Hz; $t(9)=9.44$, $p<.01$].

Differences between pitch matches from regressions based upon linear and log frequency psychometric functions did not exceed 1 Hz in the monaural condition, and 3 Hz in the illusion sequence condition. Again, this lack of difference indicates steep slopes for the individual psychometric functions, and thus the consistency of pitch judgments. If the obtained matches to the illusion sequence merely reflected subjects attending to the frequency of either component in isolation (i.e., sometimes perceiving a 400 Hz pitch and other times perceiving an 800 Hz pitch), as a reviewer suggested, then steep psychometric functions would not have been obtained. Thus, these matches appear to reflect the stable perception of pitch based upon some sort of averaging of the dichotic components.

As with all subjects, the three subjects who most accurately matched the 400 and 800 Hz tones in the monaural condition (represented by open circles in Figure 5) also gave illusion pitch matches which closely approximated the mean results from Experiment 2 (the filled circles in Figure 5). Except for the one subject who could not accurately perform the monaural pitch-matching task (open square), most subjects provided relatively accurate monaural performance. Such relatively accurate matches indicate that most subjects were capable of providing reasonable pitch matching, and all subjects provided very different pitch matches for the illusion and control conditions. Thus, the reported pitches for these subjects in the illusion condition should reasonably reflect perception for these subjects, not merely a procedural artifact

---

Insert Figure 5 about here

Experiment 4

In Experiments 1 and 2, incidence of the alternative octave illusion was not greatly affected by either ISI or sequence length (4-pairs vs 12 pairs). We thus wanted to evaluate the incidence of the illusion with shorter stimulus sequences. Experiment 4, therefore, evaluated the perception of pitch and location for a minimal illusion sequence of the dichotic 400 and 800 Hz tones (i.e., 2 dichotic pairs, and thus 1 illusion cycle).

Method

Subjects. Sixteen subjects participated in Experiment 4. Eleven were undergraduates participating for course credit. Four others were paid for their participation. The first author also served as a subject.

Procedure. There were two conditions in this experiment. In the first condition only the two-pair illusion sequences were presented, with the 800 Hz tone always in the right ear for the first pair. As in Experiment 2, the duration for each pair

was 250 ms and the interval between two pairs was 50 ms. On each trial subjects reported whether one or two tones were perceived for both the first and second dichotic pairs presented. If perceiving one pitch per dichotic pair, subjects also reported the location of the percept and which stimulus (1st or 2nd) was higher in pitch. The report procedure for this first condition was similar to that used in Experiment 1. Subjects who perceived the typical variant of the illusion subsequently participated in the second condition, which consisted of separately matching the high and low illusion pitches using the method of constant stimuli.

Results and Discussion

Thirteen subjects perceived the typical variant of the illusion. Two of the remaining 3 subjects reported hearing the identical pattern of localization, but with the low (rather than the high) pitch first. The remaining subject perceived two simultaneous pitches for each dichotic pair. Therefore, the rate of perception of the illusion pattern with a minimal illusion sequence of two dichotic pairs is as high as found for any of the longer illusion sequences studied in the earlier experiments. Once again the data showed that sequence length is not critical to the incidence of illusion.

The pitch match results for the 13 subjects who perceived high pitches toward the right ear are shown in Figure 4 as open squares. This scatter plot again indicates that the perceived pitches were distributed over a narrow range of frequencies in the lower half of the octave. The mean frequencies for perceived low and high pitches were 508 and 535 Hz (505 and 533 with log regression solution). Therefore, the mean difference in perceived pitch was 27 Hz. A t-test revealed that the difference between the matches to the perceived high and low pitches was statistically significant [$t(12)=4.19$ $p<.01$]. Although there is more variability in these results relative to the 4- and 12-tone sequences (filled symbols), the general trend in the results is quite similar.

The results of Experiment 4 indicate that the nature of pitch perception in the illusion persists even with a minimal length sequence. Although the absolute differences between high and low pitches may be somewhat smaller than found for longer (2- and 6-cycle) sequences. The results still demonstrated that sequence length is not critical to the illusion. However, we do not reject the notion that the sequence length may have effect on the perceived difference. Although not significant, the perceived pitch difference in 4- and 12-pair sequence was greater than that in 2-pair sequence. Conversely, with most subjects in Experiment 4 perceiving illusion percepts in a single cycle sequence, dichotic fusion would seem to be a contributing factor to the octave illusion, as was proposed by Bregman (1990, p 306).

Experiment 5

Experiment 4 suggested that fusion could be the critical factor in the octave illusion. If so, then subjects should still perceive a single fused tone even if only one pair of dichotic 400 and 800 Hz tones is presented. Furthermore, if pitch matches to fused tone pairs (differing in the presented location of the 400 and 800 Hz components) are similar to those found in the earlier experiments, then fusion could be argued to provide a basis for the perception of this alternative form of the octave illusion. Given the role of other stimulus and listener characteristics in fused stimuli, the central question concerning the illusion then would become why the perceived pitch toward the right ear is different (usually higher) than that toward the left ear? The two phases of the current experiment thus evaluate the nature of perception, including the possible location and pitch percepts of fused stimuli associated with a single pair of 400 and 800 Hz tones. The first phase of the experiment evaluated the number of pitches perceived and. If a single pitch was perceived, the subjects were asked to also report the location of the perceived pitch. The second phase evaluated the frequency of the perceived pitch, (assuming the perception of a single tone).

There is one major conceptual difference between pitch matching for the current and the earlier experiments. In the earlier experiments subjects heard a sequence of at least two stimuli and thus could logically be instructed to match the higher or lower pitch with the pitch of the comparison stimulus. With only a single stimulus, the relative concept of "higher" or "lower" pitch is meaningless. In this experiment we can only evaluate the pitch perceived for a specific condition of stimulus presentation (e.g., 800 Hz presented to right vs to left ear). With this inherent difference in task, the pitch matching results will be discussed in the general discussion section.

Method

Subjects. Twenty undergraduate students from SUNY-Binghamton participated as a course requirement. All subjects reported normal hearing and participated in both phases of the experiment.

Procedure. In the first phase of the experiment following presentation of a pair of dichotic 400 and 800 Hz tones, subjects pressed a button to indicate whether they perceived one or two tones. If one tone was perceived, subjects also indicated the location of the tone (left ear, left side, middle, right side, and right ear). There were 20 randomized trials in this phase of the experiment. For 10 trials, the 800 Hz tone was presented to the right ear; in the other 10 trials, the 800 Hz tone was presented to the left ear.

The second phase of the experiment was the pitch-matching task using the method of constant stimuli in which a single pair of 400 and 800 Hz pure tones was dichotically presented for 250 ms, followed by a 250 ms silence interval, then a 250 ms comparison tone. Subjects pressed one of two buttons to indicate whether the comparison tone was higher or lower in pitch than the initial tone.

Results and Discussion

Table 2 summarizes the results of the first phase of the experiment. When the 800 Hz tone was presented to the left ear, subjects reported hearing one tone on 92% of all trials. A single, fused tone was perceived either in the left ear or toward the left side of the head on 47% of trials, at a central location on 32% of trials, and either toward the right side of the head or directly in the right ear on only 13% of trials. A chi-square test confirmed that when 800 Hz was presented to the left ear, the tone is more likely to be perceived toward the left (chi^2 18.93, p< .02). When the 800 Hz tone was presented to the right ear

a single tone was perceived on 96% of trials. A singular pitch was perceived toward the right side or ear on 68% of trials, at a central location on 20% of trials, and to the left side or ear on only 8% of the trials (the remaining 4% representing dual-pitch perception). The location of the perceived tone was significantly more toward the right side when 800 Hz was presented to the right ear ($chi^2=63$, $p<.01$). Furthermore, the incidence of fusion and the pattern of perceived locations for the fused tone was quite consistent across individual subjects. The results of phase one therefore demonstrate that fusion occurs for a single dichotic pair of harmonic stimuli, and that the fused tone is more likely to be perceived toward the ear receiving the high frequency input.

Insert Table 2 about here

However, this tendency was stronger when the higher frequency was presented to the right ear, indicating that the lateralization-by-frequency effect is additionally influenced by a type of right ear dominance. We realize that this conceptualization only describes the findings, and does not provide an explanation of results. However, possible contributing factors for lateralization are discussed in more detail in the general discussion.

The pitch-match results for the 20 subjects, shown in the open circles of Figure 4, are highly similar to the results from the 2-, 4-, and 12-pair sequences from our earlier experiments. These data again are concentrated in the lower portion of the octave range, but they are distributed sufficiently to compute a reliable regression line ($r^2=.82$). The slope of this line is .87, which is close to that found with the initial 4-pair sequence (Figure 3). The slope approaching 1 and the high correlation coefficient respectively indicate that there is an approximately constant difference in frequency for the high and low pitch matches. When the right ear received the 800 Hz tone, the mean pitch was 524 Hz (521 Hz for log frequency); when the left ear received 800 Hz tone, the mean pitch was 503 Hz (501 Hz for log frequency). The perceived difference is 21 Hz. A t-test showed that the perceived difference between these means was statistically significant [$t(19)=4.38$, $p<.01$]. Thus, not only did fusion occur for a single pair of dichotically presented 400 and 800 Hz tones, but the pitch perceived when the 800 Hz tone was presented to the right ear was significantly higher than when 800 Hz was presented to the left ear. Although smaller than the 38 to 41 Hz difference for the 4- and 12-pair sequences, the difference is quite similar to the 27 Hz difference for the 2-pair sequence.

These findings provide a strong basis (i.e., the nature of fusion under different presentation conditions) for why in the octave illusion a single tone is perceived, with pitch and location changing with the change of ear presentation. Taken as a whole, Exps. 1-5 provide support for the argument that the dichotic fusion of octave-related tones is the critical factor in the octave illusion.

General Discussion

The octave illusion has three distinctive perceptual characteristics: (1) perception of one tone at a time, (2) the perceived fused tone tends to be lateralized toward the ear receiving the higher frequency input, and (3) the tone perceived to each side(right vs left) has a different pitch. We will individually address each of these three characteristics below.

Perception of one tone at a time

Subjects in Experiment 1-4 must have fused the dichotic stimuli to perceive a single pitch; if (as found with a few musically trained subjects) two simultaneous pitches had been consistently perceived, the illusion would not be possible. Taken in conjunction with the demonstration in Experiment 5 of frequent dichotic fusion for only a single pair of octave-related tones, the data reveal that the perception of one tone at a time in at least the alternative form of the octave illusion can be attributed to the dichotic fusion of the octave-related tones. The occurrence of dichotic fusion may be due to some form of special relationship between harmonically-related, and, more specifically, octave-related stimuli. McAdams (1982) has demonstrated that subjective fusion is much higher for harmonic (shared fundamental) than for inharmonic stimuli. More recently, Buell and Hafter (1991) demonstrated an inability to segregate such harmonically related stimuli even when there is significant lateral displacement of the tones.

Although there is a perceptual affinity for harmonically related tones, we have known from at least the time of Helmholtz that octaves represent the most consonant of stimulus relationships and exhibit the highest tendency for fusion. More recent, empirical findings include those of Ward (1954), who, asking listeners to adjust a tone to match a specific pitch relationship to a presented tone, found that listeners most reliably matched the octave. Ward concluded that the subjective octave should be the most stable musical pitch relationship.

A special, central nature of perception of octave-related tones has been discussed by Terhardt (1974) and has been demonstrated recently by Demany and colleagues (Demany and Semal 1988, 1990; Demany, Semal, and Carlyon 1991). Demany, et al. (1988) dichotically presented two simultaneous, sinusoidal frequency-modulated tones to listeners. Listeners were instructed to detect phase differences between the modulated tones. Demany, et al. found that the just noticeable values of phase differences were at a minimum when the tones' center frequencies differed by close to 1200 cents (one octave). These, and similar results, have led to the conclusion that octave relationships are special in terms of perception and are central in origin.

Lateralization of the fused tone

The fused tone tended to be lateralized toward the ear receiving the higher frequency input. However, might such localization be due to differences in loudness rather than frequency? There is some indication that both relative frequency and loudness are important to the location of a pitch percept with each playing relatively different roles in different stimulus settings (e.g., Cutting 1976). Deutsch (1978, 1981) showed that the lateralization to the high frequency ear even when the amplitude of the higher frequency tone is substantially lower than the amplitude of the low frequency. Efron and Yund (1974, 1975) found

that the location of fused dichotic tones tend to be located to the side receiving the louder stimulus. In contrast to these earlier studies, our two frequencies (400 and 800 Hz) were equal in intensity and thus should differ in loudness by no more than 3 phones (Fletcher & Munson, 1933). One can obtain perceptible changes in lateralization with interaural differences of one dB or less, but this is more typical for higher frequencies than were the frequencies used in the current study. Therefore, it is still an open question whether the significant lateralization effects in the current study were the result of loudness differences. Thus, although there is insufficient evidence to draw a final conclusion on whether the location of a percept is determined by relative frequency, rather than by relative loudness or by both, the preponderance of existing results tend to favor the relative frequency hypothesis. Experiment 5 also demonstrated a relatively stronger tendency to lateralize the fused tone toward the right side when the higher frequency was presented to the right ear relative to the tendency to lateralize to the left side when presented to the left ear. This finding suggests that the right ear has some advantage in lateralizing the fused tone. Further analyses of our results (below) will allow us to estimate the relative magnitude of such ear advantage.

Perceived pitch difference

The pitch-matching results of Experiments 2-5 are consistent with the notion that the auditory system performs some type of weighted averaging of octave-related dichotic inputs to determine the pitch of the fused percept.[2] Thus, we can make no specific claims concerning which aspect of pitch were studying. Furthermore, the pitches we measured in the current study seemed to be relatively consistent across subjects. The frequency of the fused pitch, F_p, then might be described by the simple formula:

$$F_p = R*F_r + L*F_l \qquad (1)$$

In this formula R and L are relative weighting factors for the stimuli presented to the right and left ears; these weights are assumed to always sum to unity (i.e., R = 1 - L). F_r and F_l respectively are the frequencies presented to the right and left ears; in the current study F_r and F_l were either 800 or 400 Hz. If the perceived pitch was determined solely by the input to the dominant ear, then the value of R in Eq. 1 must be 1. Clearly it is not the case here for the form of the illusion studied.

Our results demonstrate a very strong, but not universal, tendency for subjects to hear a relatively higher pitch when the higher frequency (800 Hz) was presented to the right ear for sequences of 2, 4, and 12 dichotic pairs. It thus seemed more reasonable to initially focus the formula on relative frequency, rather than on ear of presentation. To accomplish this change of focus, Eq. 2 is analogous to Eq. 1, but with the modification to reflect weights for the higher and lower frequency tones.

$$F_p = H*F_h + L*F_l \qquad (2)$$

F_h and F_l again are 800 and 400 Hz (now independent of ear of presentation); H and L are the respective weights for 800 and 400 Hz tones and again must sum to unity. Since Eq. 2 assumes the perceived pitch to be a function of the presented location of the 800 Hz tone, we must solve Eq. 2 using a different pair of high and low frequency weights for each of these two presentation conditions.[3] The two pairs of weights should therefore reflect the perceived higher and lower pitches when 800 Hz was presented to either the right or left ear, with the difference in weights reflecting the contribution of ear advantage.

Insert Table 3 about here

Table 3 summarizes the H and L weights from Eq. 2 as a function of sequence length and perceived pitch. The weights H and L are consistent across different sequence lengths: the means for H and L respectively were 0.36 and 0.64 when the higher pitch was perceived, and 0.26 and 0.74 when the lower pitch was perceived. These weights provide quantification that the perceived pitch is a product of combination of frequencies presented to two ears.

The 1-pair pitch match results (Experiment 5) were based solely on presentation condition and not on relative pitch. Experiment 5 indicated that there was a strong, but not absolute tendency for the location of a fused tone to be correlated with the ear receiving the higher frequency tone as reported by Deutsch (1978, 1981). If these measurements of localization tendency reflect similar tendencies for underlying pitch process (e.g., if perceived pitch and location are highly correlated), then we should be able to use the results summarized in Table 3 to predict the pitch-matches for the one pair condition. Our new formula substitutes the mean values of H and L (from Table 3) into Eq. 2 for each of the two perceived pitches. The portion of the formula for each pitch then is weighted by the probability of localization toward each ear. The resulting formula is

$$F_p = P_r*[.36*F_h + .64*F_l] + P_l*[.26*F_h + .74*F_l]. \qquad (3)$$

where P_r and P_l are the relative probability of localization toward each ear, as determined in Experiment 5. Eq. 3 must be applied separately for each of the two presentation conditions (800 Hz to right or left ear). The values of F_p are 513 and 540 Hz (a 27 Hz difference). Both actual mean pitch matches were about 10 Hz higher than predicted. However, in this variant of octave illusion, listeners perceived consistent difference in frequency between the two pitches, but vary in the absolute values of the pitches. Thus, the difference in pitch may be more important than the individual frequencies of either the perceived high or low pitches. The observed frequency difference of 24 Hz is sufficiently similar to the predicted 27 Hz difference to suggest that fusion of a single pair of octave-related tones reflects the weighted combination of inputs to both ears and that fusion seems to be the most critical contributing factor to the octave illusion.

The weights in Table 3 also reflect the relative contribution of fusion and ear dominance. The weight H was 0.36 when 800 Hz was presented to the right ear and 0.26 when 800 Hz was presented to the left ear. Thus, the input to the right ear always is weighed more heavily by 0.1 than the input to the left ear. This weighting can be used as the basis of a more detailed quantification to account for the frequencies of the perceived pitches. Because the weights sum to unity, we can convert the weights to percentages. Average perceived pitch can be represented as the sum of 26% of the higher frequency and 74% of the lower frequency and 10% of the frequency presented to the right ear. For example, when 800 Hz tone was presented to the right ear and 400 Hz tone was to the left, the perceived pitch in Hz predicted using the equation

$$F_p = .26*800 + .64*400 + .10*800, \quad (4)$$

yielding a predicted F_p of 544 Hz. Using the same equation, F_p is 504 Hz when 400 Hz tone is presented to the right ear (with a predicted 40 Hz difference). The actual means of the perceived pitches of 543 and 504 Hz (39 Hz difference) are consistent with these computed frequencies and are independent of sequence length.

The estimated weights (H and L) are important not in terms of their specific values, but rather in terms of the processing they reveal. Weighted averaging of inputs is consistent with a large literature that suggests a central mechanism which is sensitive to octave-related stimuli (e.g., Demany & Semal 1988, 1990, Demany, Semal & Carlyon 1991). Ear difference in tonal input weighting also has been reported. For example, Ward (1954) found that two ears of a single observer gave different pitch match results with one ear consistently giving relatively higher pitch matches for a specific tone (e.g., binaural diplacusis).

Existing pitch models

The nature of fused pitches in the illusion cannot be easily explained by models of pitch perception of complex tones; the focus of these models is the explanation of the residue and other pitch percepts in complex stimuli where pitch does not correspond to place along the basilar membrane. For reasons to be discussed below, we feel that the pitch percepts studied in the current research represent a very different type and level of processing than that addressed by such models. There are two broad classes of such modern models concerning complex pitch perception: Pattern Recognition and Temporal models. Pattern recognition models assume that the pitch of a complex tone is based upon neural signals corresponding to primary sensation, e.g., the pitches of the individual partials. Goldstein's optimum processor theory (1973), Terhardt's (1972a,b, 1974) pitch perception theory, and Houtsma and Goldstein's central pitch processing theory belong to this group. Goldstein's model (1973) predicts that the pitch of a complex tone corresponds to the 'best fit' of the harmonic series in the complex tone while Terhart's (1974) model suggests that the perceived pitch of a complex tone would always be a subharmonic of a dominant partial (resolvable partials), rather than the lowest partial. We are describing a single pitch percept for stimuli which should be resolvable (in terms of critical band differences) even if presented to one ear and whose partials should be perfectly coincident.

Temporal models assume that the pitch of a complex tone is based upon the time interval between corresponding points in the fine structure of the signal close to adjacent enve' e maxima (Schouten, Ritsma, & Cardozo, 1962; Wightman, 1973). Schouten's theory suggested that the pitch of a comple.. tone corresponds to the *most prominent component in* that sound. Wightman's pattern-transformation theory was not aimed to predict pitch-match data. However, because one of our stimuli is the harmonic of the other, there is no simple, predictable fine temporal structure among potential particles which does not correspond to the temporal properties of one of the original stimuli. However, the pitch models all were developed to address very different concerns related to pitch perception. Furthermore, the uniqueness of the octave relationship is probably providing the perceptual system with an unusual, and most-likely modified sensory information. It would be interesting to evaluate the pitch of dichotic, octave-related stimuli which are above 5 kHz where most models, and evidence, indicate that the coding of pitch is known to operate on a different basis.

Individual differences in perception

As reported by Deutsch (1974, 1980, 1983a), we consistently observed individual differences in the perception of the reported variant of octave illusion, with many listeners usually perceiving the illusion, but with some listeners seldom perceiving the illusion. Furthermore, we typically found that extent of musical experience seemed to be negatively correlated with perception of the specific variant of the illusion.

The difference in perceptual tendencies between musicians and nonmusicians in the illusion may be due to the nature of musical training. Musicians are trained to listen to music in an "analytic" fashion, with many trained to be able to recognize both tone location and what instruments were playing them. This training may account for why musicians often are aware of two simultaneous tones in an illusion sequence. The notion that musical training produces behavioral differences that have been well documented (e.g., Helmholtz, 1863; Houtsma and Goldstein, 1972; Houtsma, 1979; Cross and Lane, 1963). Helmholtz (1863) reported that complex periodic sounds can be perceived "synthetically" or "analytically" (i.e., perceived as one sound or in terms of individual partials). Cross and Lane (1963) reported that listening "synthetically" and "analytically" can be controlled by previous training, and Houtsma (1979) has suggested that musically experienced listeners have a much stronger tendency to perceive complex sounds analytically than musically naive listeners. Although not addressed in the current research, it also is possible that such an "analytic-holistic" distinction may reflect either differences in hemispheric dominance (e.g., Deutsch 1982) or, alternatively, differences in the efficiency of encoding component stimuli.

In contrast to our musically trained subjects, results obtained with musically naive listeners indicate a holistic percept in which pitch seems to reflect a weighted averaging of both components. Clearly, such an averaging of dichotic information must reveal processing that is occurring at a more central level, rather than at (or directly derived from the operation of) the sensory mechanism. We note that musicians also often report perceiving an additional fused percept shifting in localization, which probably is the output from a central, octave-related mechanism responsible for fusion. Existing pitch models are based directly on the output of the latter (peripheral) type of processing, and thus are not designed to address such a weighting of information for a global percept.

We thus claim that our results are addressing a different type and level of processing than a century of unified research on the manner in which listeners normally perceive pitch. In fact, we believe that our findings, and those of Deutsch (1974a) on the common variant of the octave illusion, both probably reflect an inability to segregate and evaluate component stimuli, rather than a general difficulty in accurately encoding frequency at the sensory level. Indeed, most of our musically naive subjects could accurately match monaural stimuli, but still matched illusion pitches as intermediate between Left and Right component

We thus believe that we have been investigating an attentional limit based on the analytic-holistic distinction. Additional support for such notion comes from the fact that musicians also reported a type of holistic percept, which shifted in lateralization with the alternation of stimuli. Although it may be difficult for musicians to map pitch for this weak holistic percept (having to additionally ignore the strong pitches of perceptually isolated components), it is conceivable that the holistic percept, in fact, may be similar in nature to the illusory percept commonly reported by our musically naive subjects. Thus, the variant of the illusion also could potentially reflect frequency averaging in musicians, and merits some attempt at further research. Currently, however, regardless of the types of processing which are reflected by individual differences in the variant of the octave illusion, it is clear that existing evidence indicates that musical experience is negatively correlated with perception of this variant of the illusion.

Summary

The current study indicates that the critical contributing factor to the octave illusion is dichotic fusion, which provides the basis for the perception of one tone at a time. A secondary contributing factor is a right ear advantage for weighing input; it is the ear advantage which contributes to the slight shift in pitch for this variant of the illusion.

The underlying processing of dichotic pairs of octave-related stimuli in terms of the perception of pitch and location is not easily explained by current models of pitch perception whose goals and focus are not on such simple stimuli. As has been conjectured for pitch, there does appear to be a central mechanism responsible for frequency averaging of octave-related dichotic tones for the current listeners. However, it is not obvious whether such a mechanism is restricted to octave-related stimuli. Since fusion, the primary contributing factor to this illusion, is not limited to octaves, it is possible that the illusion could be produced under a wide variety of stimulus conditions. In a future submission, we also will present data which demonstrates a similar illusion based upon frequency components that are not octave-related.

References

Akerboom, S., Hoopen, G., & Knoop, A. (1985). Does the octave illusion evoke the interaural tempo illusion? Perception & Psychophysics, 38(3), 281-285.

Bregman, A.S. (1990) Auditory Scene Analysis. The MIT press, Cambridge, Massachusetts.

Cross, D., and Lane, H. (1963). Attention to single stimulus properties in the identification of complex tones. Experimental analysis of the control of speech production and perception, University of Michigan ORA Rep. No 05613-1-p.

Cutting, J. (1976). Auditory and linguistic processes in speech perception: inference from six fusions in dichotic listening. Psychological Review, 83(2), 114-140.

Demany, L. and Semal, C. (1988). Dichotic fusion of two tones one octave apart: Evidence for internal octave templates. Journal of the Acoustical Society of America, 83, 687-695.

Demany, L. and Semal, C. (1990). Harmonic and melodic octave templates. Journal of the Acoustical Society of America, 88, 2126-2135.

Demany, L., Semal, C., and Carlyon, R. (1991). On the perceptual limits of octave harmony and their origin. Journal of the Acoustical Society of America, 90, 3019-3027.

Deutsch, D. (1974a). An auditory illusion. Journal of the Acoustical Society of America, 55, S18-S19.

Deutsch, D. (1974b). An auditory illusion. Nature, 251, 307-309.

Deutsch, D. (1975a). Musical illusion. Scientific American, 233, 92-104.

Deutsch, D. (1975b). Two-channel listening to musical scales. Journal of the Acoustical Society of America, 57, 1156-1160.

Deutsch, D. (1976). Lateralization by frequency in dichotic tonal sequence as a function of interaural amplitude and time difference. Journal of the Acoustical Society of America, 60, S50(a).

Deutsch, D. (1978a). Binaural integration of tonal patterns. Journal of the Acoustical Society of America, 64, S146(a)

Deutsch, D. (1978b). Lateralization by frequency for repeating sequence of 400-Hz and 800-Hz tones. Journal of the Acoustical Society of America, 63, 184-186.

Deutsch, D. The octave illusion and auditory perceptual integration (1981). In J.V. Tobias and E.D. Schubert (Eds.). Hearing Research and Theory (Volume 1). Academic Press: New York.

Deutsch, D., & Roll, P.L. (1976). Separate "what" and "where" decision mechanisms in processing a dichotic tonal sequence. Journal of Experimental Psychology: Human Perception and Performance, 2, 23-29.

Deutsch, D. (1988). Lateralization and sequential relationships in the octave illusion. Journal of the Acoustical Society of America, 83(1), 365-369.

Efron, R., and Yund, E.W. (1974) Dichotic competition of simultaneous tone bursts of different frequency: I. Dissociation of pitch from lateralization and loudness, Neuropsychologia, 12, 249-256.

Efron, R., and Yund, E.W. (1975) Dichotic competition of simultaneous tone bursts of different frequency. III. The effect of stimulus parameters on suppression and ear dominance function. Neuropsychologia, 13, 151-161.

Fletcher, H. and Munson, W.A. (1933) Loudness, its definition, measurement and calculation. Journal of the Acoustical Society of America, 5, 82-108.

Goldstein, J.L. (1973). An optimum processor theory for the central formation of the pitch of complex tones. Journal of the Acoustical Society of America, 6, 1496-1516.

Houtsma, A. J. M. (1979). Musical pitch of two-tone complexes and predictions by modern pitch theories. Journal of the Acoustical Society of America, 66, 87-99.

Houtsma, A. J. M., and Goldstein, J. L. (1972). The central origin of the pitch of complex tones: Evidence from musical interval recognition. Journal of the Acoustical Society of America, 51(2), 520-529.

Schouten, J.F., Ritsma, R.J., and Cardozo, B.L. (1962). Pitch of the residue. Journal of Acoustical Society of America, 34, 1418-1424.

Terhardt, E. (1972a). Zur Tonhohenwahrenhmung von Klangen. I. Psychoakustische Grundlagen. Acustica 26, 173-186.

Terhardt, E. (1972b). Zur Tonhohenwahrenhmung von Klangen. II. Ein Funktionsschema. Acustica 26, 187-199.

Terhardt, E. (1974). Pitch, consonance and harmony. Journal of Acoustical Society of America, 55, 1061-1069.

Von Helmholtz, H. L. F. (1863). Die lehre von den tonempfindungen als physiologische Grundlage für die theorie der musik (F. Vieweg & Sohn, Braunschweig).

Ward, W.D. (1954). Subjective musical pitch. Journal of The Acoustical Society of America, 26, 369-380.

Wightman, F.L. (1973). The pattern-transformation model of pitch. Journal of The Acoustical Society of America, 54, 407-416.

Acknowledgments

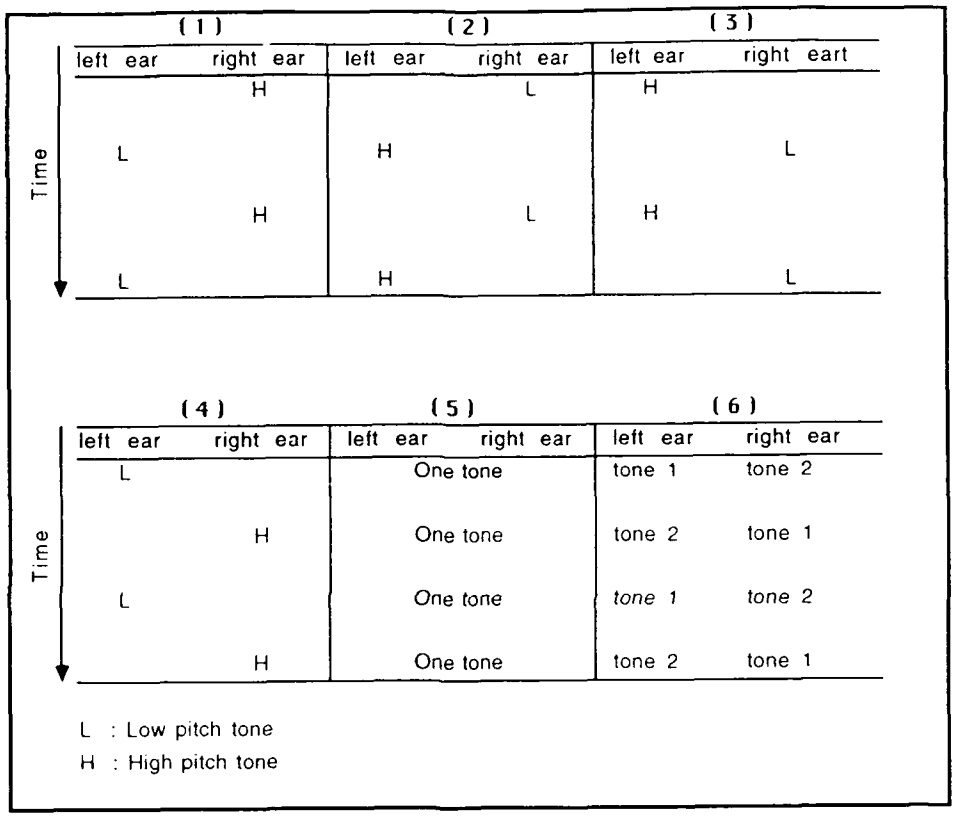
Based upon work supported by National Science Foundation Grant BNS8911456 and Grant F496209310033 from the Air Force Office of Scientific Research. Opinions, findings, conclusions, and recommendations are the authors' and do not necessarily reflect views of the granting agencies.

Endnotes

1. Highly practiced musicians tended to hear two simultaneous complex sounds which were toward, but not in, each ear and whose pitches, while different from each other, seemed to be intermediate to the physical stimulus.

2. Although there are several different pitch-related percepts, our purpose was to study the pitch changes associated with perception of the illusion, and we did not make any attempt to explain or explicitly instruct the subjects to respond to any specific aspect of perceived pitch.

3. In a sequence of stimuli we cannot perfectly correlate a specific perception with the presentation of stimuli to a specific ear, but can determine whether the subject is responding to the higher or lower pitch, and we do know that such pitch perception is highly stable with individual subjects. We therefore must apply the formula based upon perceived pitch, assuming a difference in the H and L weights for the two percepts. We are faced with the opposite problem when we present a single pair (half cycle) of stimuli; here we know what was presented on each trial, but not which pitch was perceived—we only know the overall pitch average for each presentation condition



Time	(1) left ear	(1) right ear	(2) left ear	(2) right ear	(3) left ear	(3) right eart
		H		L	H	
	L		H			L
		H		L	H	
	L		H			L

Time	(4) left ear	(4) right ear	(5) left ear	(5) right ear	(6) left ear	(6) right ear
	L		One tone		tone 1	tone 2
		H	One tone		tone 2	tone 1
	L		One tone		tone 1	tone 2
		H	One tone		tone 2	tone 1

L : Low pitch tone
H : High pitch tone

Table 1: The six perceptual patterns given to Exp.1 subjects defining distinct responses.

Reported Perception	800 Hz to left ear	800 Hz to right ear
2-tones	8	4
1-tone	92	96
1-pitch locations		
Left ear	17	2
Left side	30	6
center	32	20
Right side	8	28
Right ear	5	40

Table 2: Number of perceived tones and location of fused tone given a single (800/400Hz) dichotic stimulus pair (expressed as percentage of trials).

	Perceived High pitch			Perceived Low pitch		
Sequence Length	Freq.	H	L	Freq.	H	L
2-pair	535	.34	.66	508	.28	.72
4-pair (levitt)	550	.38	.62	501	.25	.75
4-pair	548	.37	.63	510	.28	.72
12-pair	538	.35	.65	497	.24	.76
Mean	543	.36	.64	504	.26	.74

Table 3 : Weighing factors (H/L) in different sequence length.

Figure Captions

Figure 1. The original stimulus configuration and most typical reported perception of the octave illusion from Deutsch (1974a).

Figure 2. Stimulus configuration and typical illusory perception for Experiment 1 of the current study.

Figure 3. Pitch-matching results in Experiment 2 using the Levitt up-down procedure. The regression line ($y = .91x + 95.3$) indicates the consistent perceived pitch difference.

Figure 4. Pitch-matching results for different sequence lengths in Experiments 2, 4, and 5 using the method of constant stimuli. The filled circles and filled squares respectively represent individual pitch matches for 4-pair and 12-pair sequences in Experiment 2. The open squares represent pitch matches for 2-pair sequences in Experiment 4. Finally, the open triangles represent pitch matches for 1 dichotic pair of tones in Experiment 5; the regression line is computed only for data from this condition.

Figure 5. Individual pitch-matching results for monaural sequences of alternating 800 and 400 Hz tones (open symbols), as well as for 4-pair illusion sequences (filled symbols). Circles represent data from the 3 subjects who most accurately matched monaural stimuli. The square represents the data from 1 subject who could not accurately match monaural stimuli. The triangles represent data from the remaining 6 subjects, who matched monaural stimuli with moderate accuracy. The matches to the illusion sequence are significantly different from the monaural sequences, and are similar to the mean results of Experiment 2 for the same sequence length (the dark asterisk).

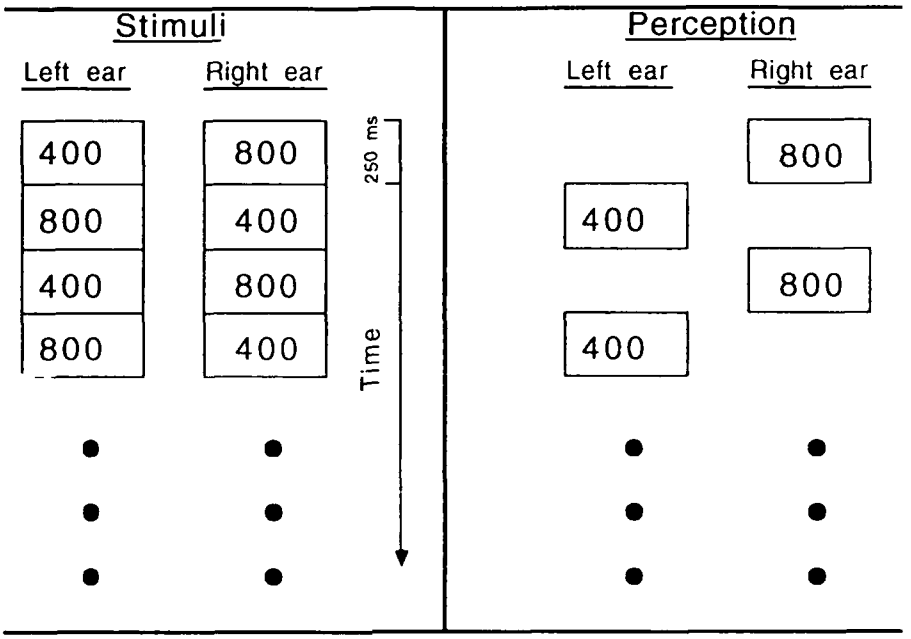

Stimuli
Left ear
Right ear
400
800
400
800
800
400
800
400
250 ms
Time
Perception
Left ear
Right ear
400
400
800
800

Figure 1

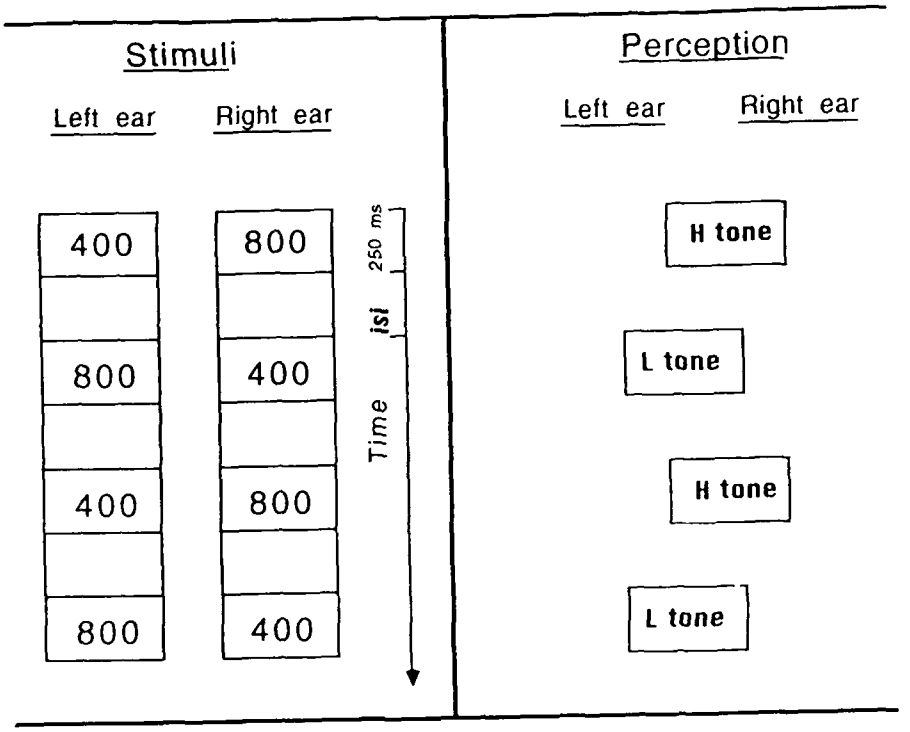

Stimuli
Left ear
Right ear
400
800
800
400
400
800
800
400
250 ms
isi
Time
Perception
Left ear
Right ear
H tone
L tone
H tone
L tone

Figure 2

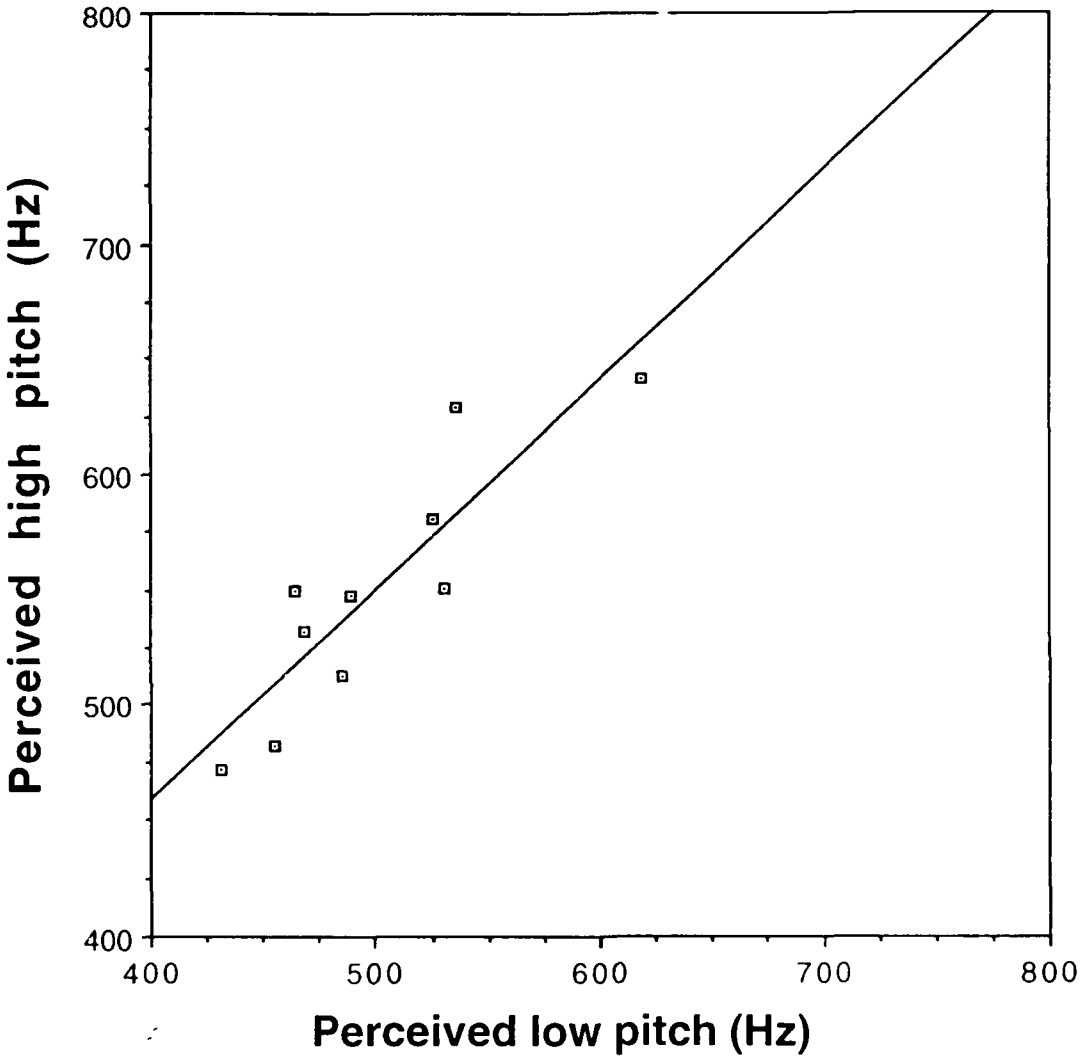

800
700
600
500
400
400
500
600
700
800
Perceived high pitch (Hz)
Perceived low pitch (Hz)

Figure 3

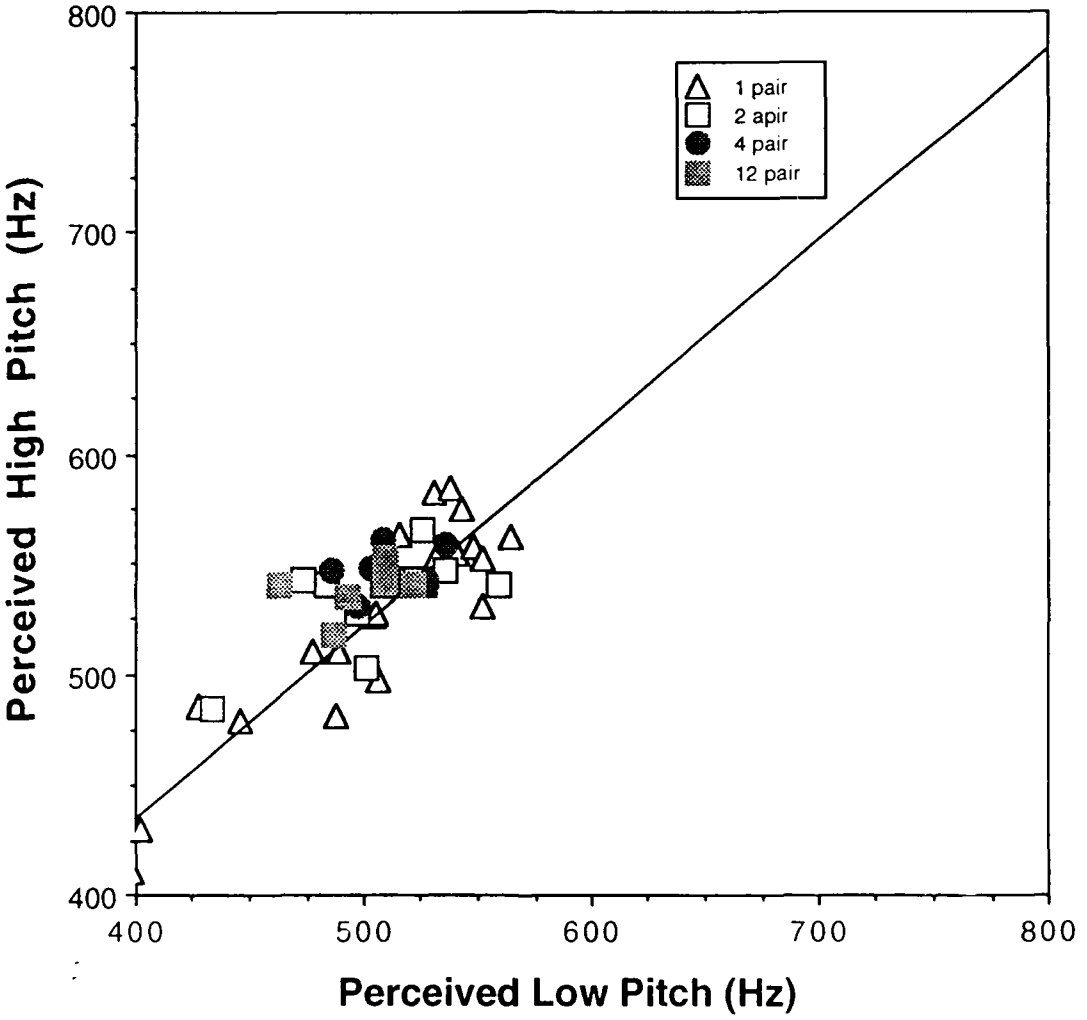

800
700
600
500
400
400
500
600
700
800
Perceived High Pitch (Hz)
Perceived Low Pitch (Hz)
1 pair
2 apir
4 pair
12 pair

Figure 4

Method of Constant Stimuli Replication

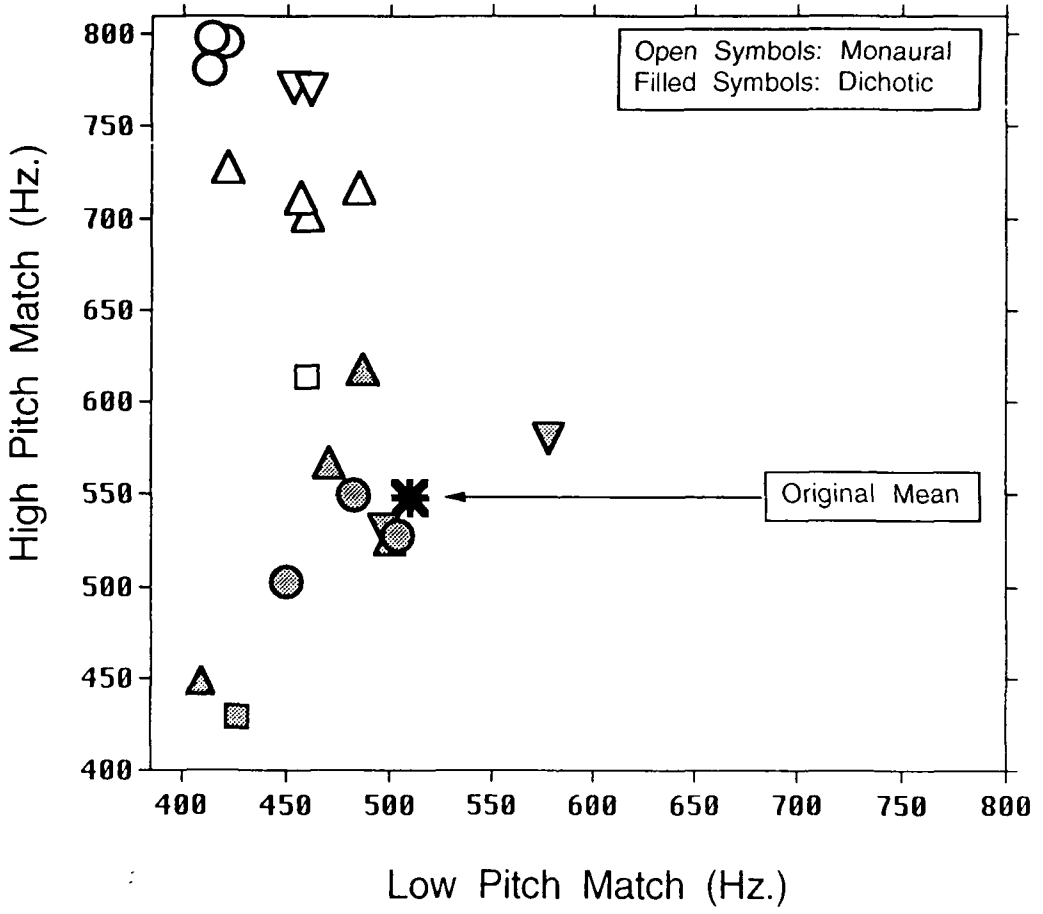


Figure 5

An Auditory Analogue to Feature Integration*

Michael D. Hall and Richard E. Pastore

Psychoacoustics Laboratory, Department of Psychology
State University of New York at Binghamton
Binghamton, NY 13902-6000

Two experiments were conducted to establish feature integration for audition using search tasks analogous to those typically used in vision. Arrays of varying complexity (# of items) were constructed using musical tones distinguished by their pitch, timbre, and location. Subjects searched arrays for cued pitches timbres, or conjunctions of both.

Subjects detected with high confidence the incorrect conjunction of pitch and timbre presented only separately. Also, conjunction search latencies increased, and accuracy decreased, with increased array complexity. Following the logic from visual research, these findings reflect the focusing of analogue attention to conjoin features.

A similar pattern of results was obtained for searches focusing on only pitch or timbre. Thus, pitch and timbre may be feature conjunctions. However, given extensive experience with the stimuli, instrument timbres may be processed more like distinct features.

Introduction

Theory

A number of generalizable models of attention recently have come from vision research. One such model is Feature Integration Theory, or FIT (e.g., Treisman & Gelade, 1980). According to FIT, attention is initially distributed, and the observer abstracts in parallel all individual features and overlearned (automatized) feature conjunctions. Analogue attention is then focused to integrate features at a location, and thus perceive objects.

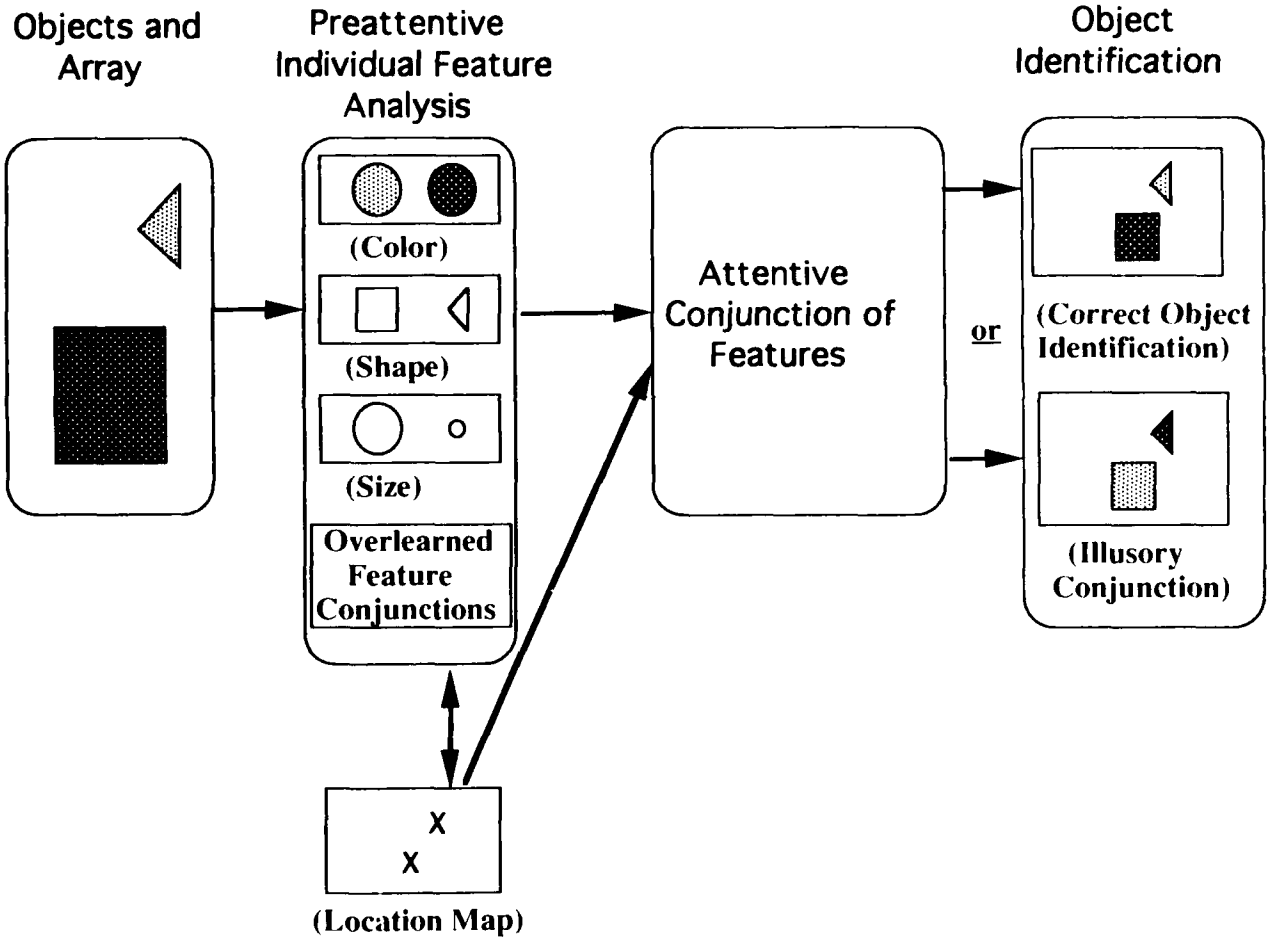
Original evidence for FIT comes from visual search tasks, where S's search for single features or feature conjunctions in arrays of varying complexity. Whereas single-feature search times are relatively unaffected by array complexity, conjunction search times linearly increase with increasing array complexity, presumably reflecting focused attention. Also, when attention becomes overloaded, as when many items are presented, illusory conjunctions (wrong, but confidently perceived combinations of presented features) occur often.

Possible Auditory Applications

Although illusory conjunctions have been conjectured to occur in audition, such as with the mislocalization of components of dichotic musical chord stimuli (e.g., Hall & Pastore, 1992), no direct test of FIT has been made for audition. The current study therefore used musical stimuli and methods analogous to visual search tasks in an evaluation of the applicability of FIT to audition. Experiment 1 examined search performance as a function of array complexity both for assumed single features and conjunctions of those features. Conditions were included where illusory conjunctions were evaluated. Experiment 2 then attempted to verify suggestions of parallel search for single features by examining the effects of experience with the stimuli.

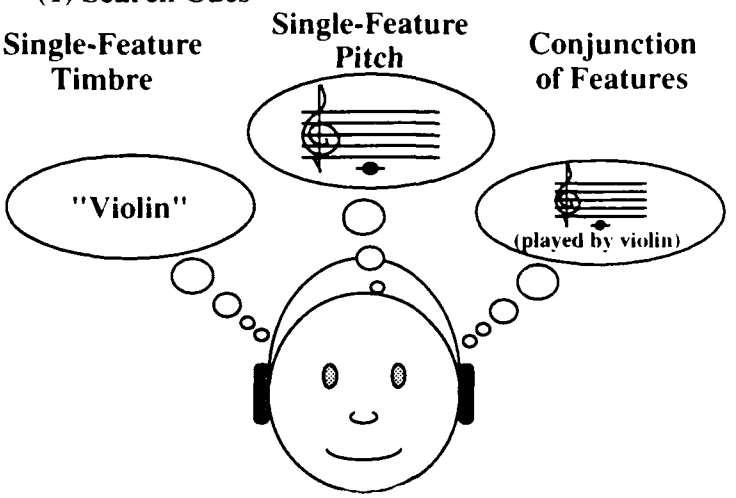
* Poster presented at the 34th annual meeting of the Psychonomic Society, Washington, DC, November 5, 1993

Feature Integration Theory



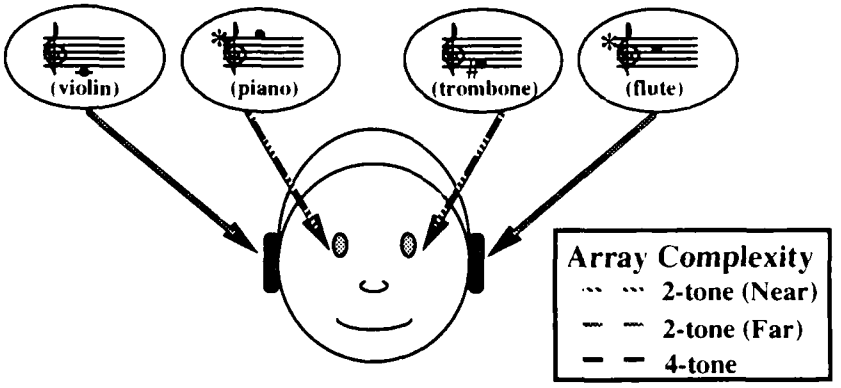
General Method

(1) Search Cues



(2) Cues and Arrays = 2.0 s each; ISI = .5 s

(3) Array Presentation



* Indicates flat mistuning from indicated pitch.

(4) Question: "Was the cued target in array?"

(a) "Yes" or "No" Response

(b) Confidence Rating (conjunction trials only)

1	2	3	4	5	6	7
confident yes			unsure			confident no

(5) Nature of Search Cues

	Cues in Array	Conjunction?
Single-Feature Search		
Valid	1	--
Invalid	0	--
Conjunction Search**		
Valid	2	Yes
Invalid(+)	2	No*
Invalid(-)	0	--

* Confident "yes" response = illusory conjunction.

**Experiment 1 only.

Cues:

--"Single-feature" (timbre or pitch) or "feature conjunction" search.

--Timbre cues = 4 visual char. for instrument: violin, piano, trombone, clarinet, or flute.

--*Pitch cues = sine tones: 262, 370, 509, 762, or 1078 Hz.*

--Conjunction cues = pitch by natural timbre.

--Cues: "Valid" (in array) or "Invalid" (not in array); p = 0.5

--Features of invalid conjunction cues:
"Invalid(+)" (present, not conjoined); p = 0.5
"Invalid(-)" (none in array); p = 0.5

--Confident positive response for Invalid(+) cue = illusory conjunction.

Arrays:

--Varied complexity (# of tones) and tone distance (near vs. far).

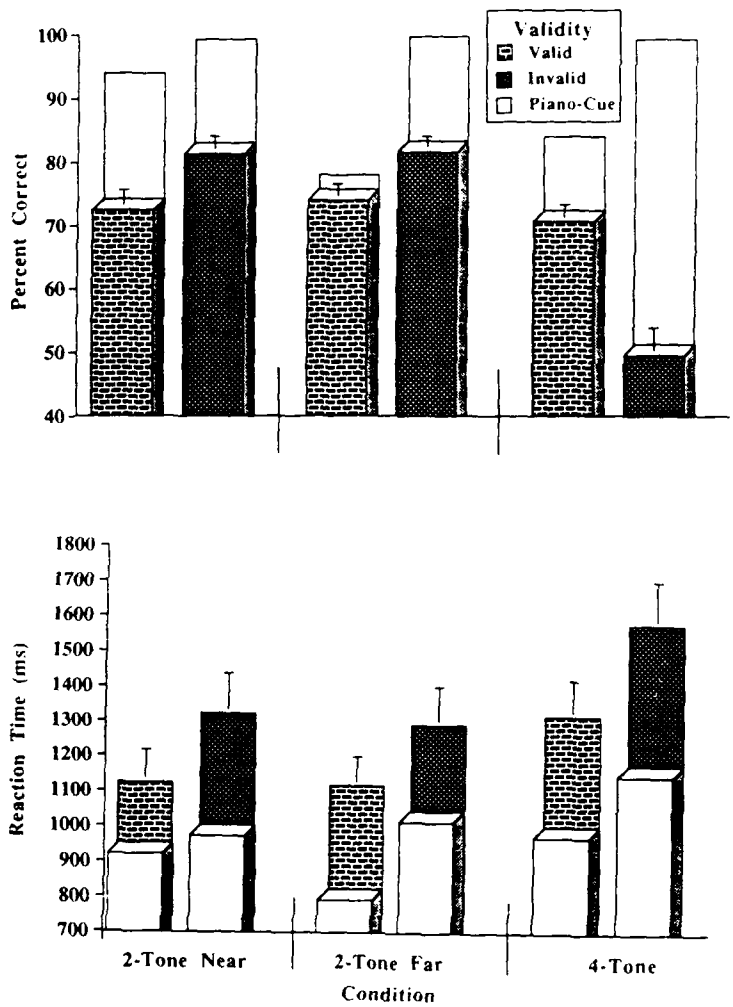
--Unique tone localization produced by:

(1) manipulating interaural time disparities,

(2) inharmonic pitches separated by ≈ 1/2 octave, with distinct timbres (see above).

Experiment 1: Auditory Search

Single-Feature Search for Timbre



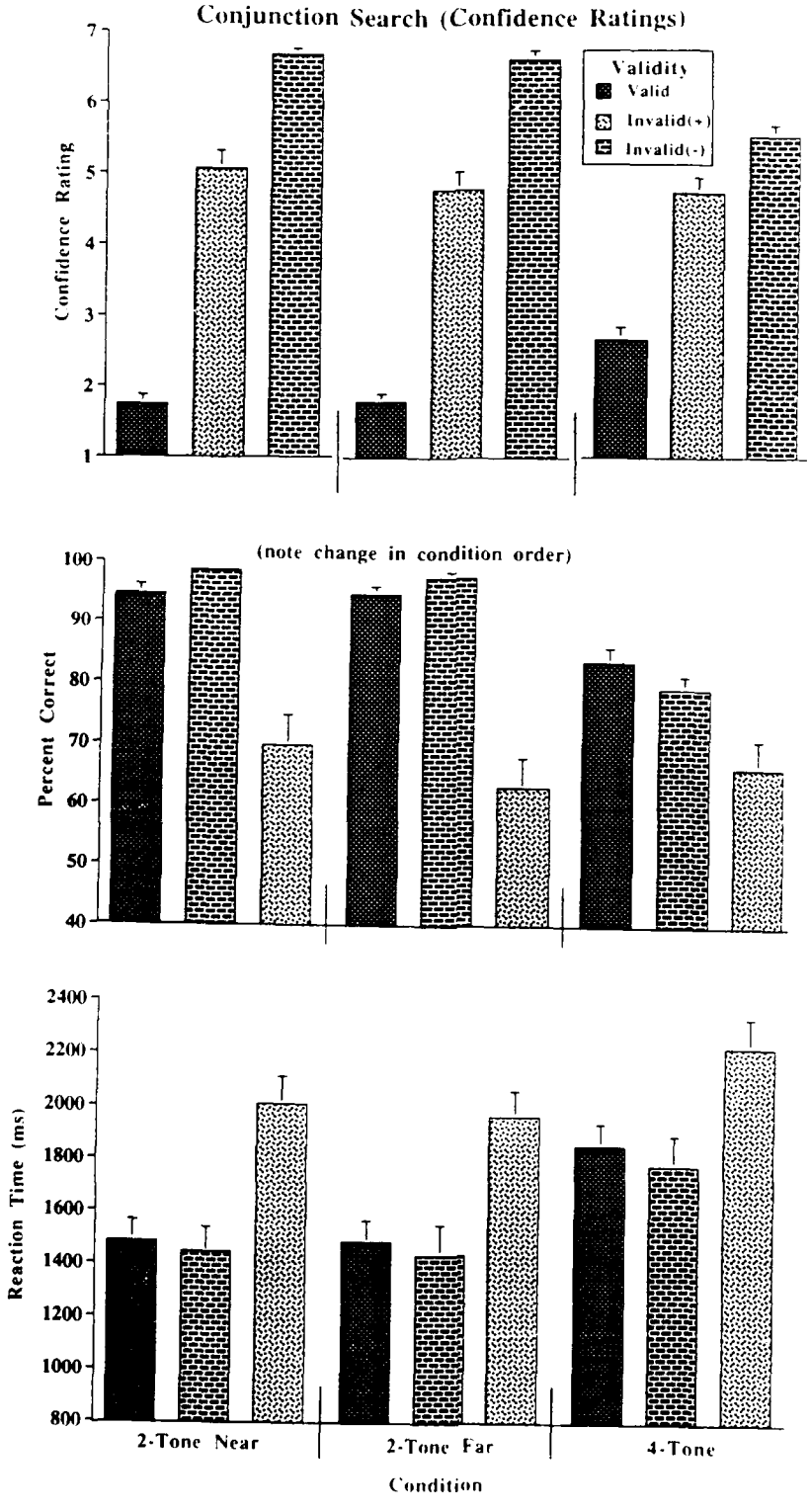
Timbre results displayed, 10 S's (Analogous results for pitch)

--Significantly decreased accuracy, and increased RT, with increased array complexity (2- vs. 4-tone); ***not consistent with (feature) parallel search.***

--Significantly faster responses to valid cues; ***typical of search results.***

--No distance effect (2-tone Near vs. Far).

--Since many S's were pianists, piano timbre may have acted as a single (overlearned or acquired) feature. Accuracy increased, and RT decreased (***thus better approximating parallel search***), but still some indication of array complexity effects.


Conjunction Search (Confidence Ratings)
Confidence Rating
7
6
5
4
3
2
1
Validity
Valid
Invalid(+)
Invalid(-)
(note change in condition order)
Percent Correct
100
90
80
70
60
50
40
Reaction Time (ms)
2400
2200
2000
1800
1600
1400
1200
1000
800
2-Tone Near
2-Tone Far
4-Tone
Condition

Illusory Conjunctions

Mean Error Rates = illusory conjunctions + general errors		
Audition: Hall & Pastore ('93) (minus timbre confusions)		Vision: Treisman & Schmidt ('82)
Near (2-Tone)	.18	.23
Far (2-Tone)	.25	.16
4-Tone	.24	
High Confidence Error Rates (minimizes general errors)		
Near (2-Tone)	.12	
Far (2-Tone)	.17	
4-Tone	.14	
Mean	.14	.15

Feature Integration Evidence: Displayed for confidence ratings, 10 S's
(Similar results for "*Yes/No*" *task, but with faster* RTs)

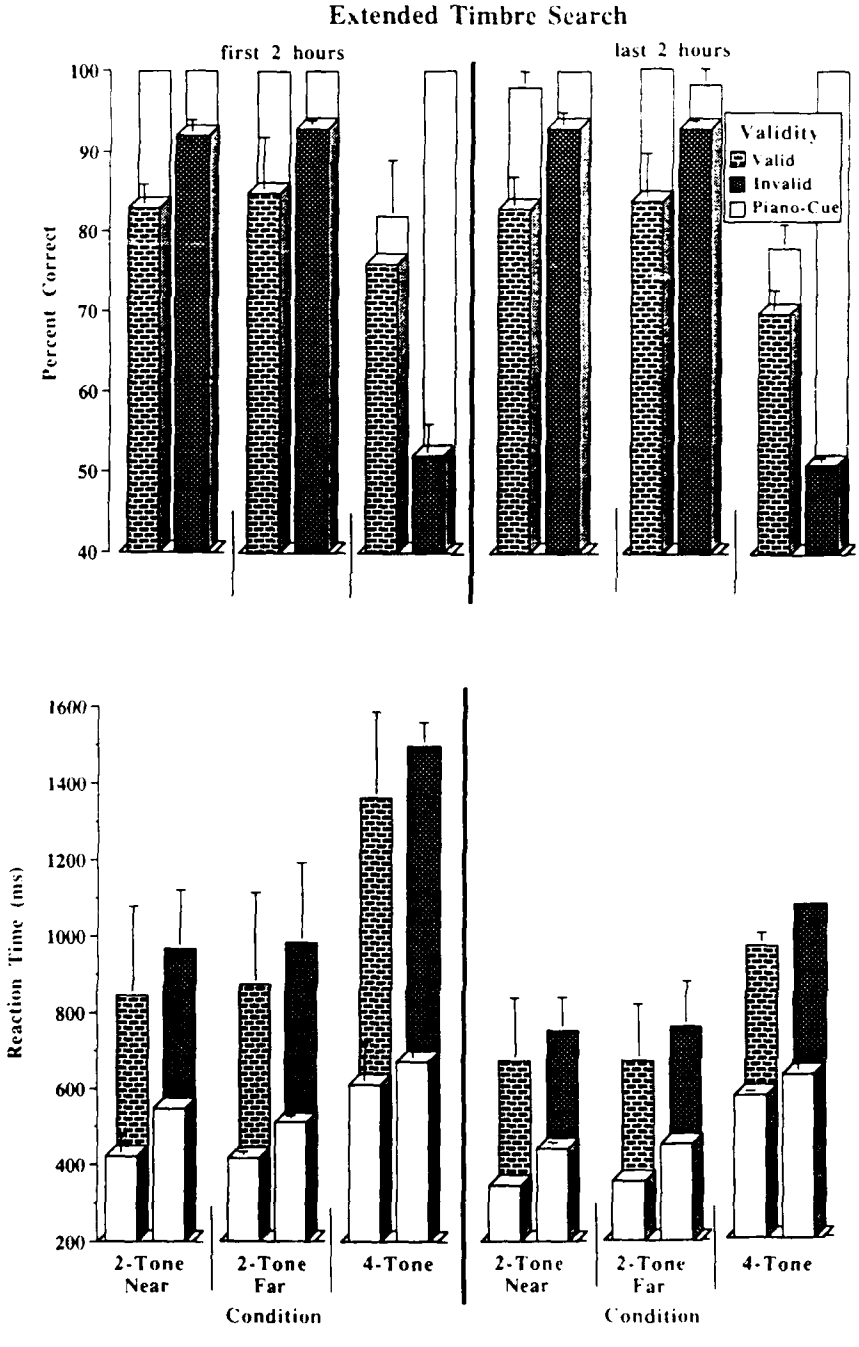
--Significantly less confident*, less accurate*, and slower ratings with increased array complexity (2-vs. 4-tone); ***consistent with predicted attention-demanding serial search***

--Significantly fewer high confidence responses, more errors, and slower RT on Invalid(+) cue trials ***(see below)***

--Moderate rates of errors where subjects responded with high confidence on Invalid(+) cue trials; ***strong evidence of illusory conjunctions*** Estimated rates comparable to rates in vision

* Marginal simple effects for Invalid(+) cue trials.

Experiment 2: Experience Effects



Displayed for 2 pianists (Similar tendency for other 2 S's)

Accuracy:

--Significantly decreased overall accuracy with increased array complexity (marginal for piano cue trials).

--No practice effects.

--Reduced difference for piano cue trials ***(ceiling effect?)***

RT:

--Significantly increased overall RT with increased array complexity, consistent with serial search. Also, significantly faster on valid cue trials ***(i.e., self-terminating search)***

--Marginally significant reduction of RT with practice.

--Reduced effects of array complexity given more practice, especially on piano cue trials. ***revealing trend toward automaticity***

Conclusions & Future Directions

(1) Evidence for serial search and illusory conjunctions argue for feature integration in audition.

(2) Timbre and pitch each appear to represent conjunctions of primitive features.

(3) With extensive experience, a timbre (e.g., piano) can be processed more automatically (i.e., akin to a single-feature).

(4) The single-feature search task can be used to define auditory (e.g., timbral) features.

References

Hall, M. D. & Pastore, R. E. (1992). Effects of base complexity in musical duplex perception. Proceedings of the 123rd Meeting of the Acoustical Society of America, 91(4, pt. 2), 2339 (abstract 2SP7).

Snodgrass, J. G. & Townsend, J. T. (1980). Comparing parallel and serial models: Theory and implementation. Journal of Experimental Psychology: Human Perception and Performance, 6(2), 330-354.

Treisman, A. (1982). Perceptual grouping and attention in visual search for features and for objects. Journal of Experimental Psychology: Human Perception and Performance, 8, 194-214.

Treisman, A. (1992). Perceiving and re-perceiving objects. American Psychologist, 47(7), 862-875.

Treisman, A. & Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97-136.

Treisman, A. & Schmidt, H. (1982). Illusory conjunctions in the perception of objects. Cognitive Psychology, 14, 107-141.

Work supported by AFOSR grants F496209310033 and F496209310327, and NSF grant BNS8911456

Implicit assumptions in modeling higher level auditory processes

Richard E. Pastore
Center for Cognitive & Psycholinguistic Sciences
Binghamton University
Binghamton, New York 13902-6000

Abstract

There has been growing interest in the investigation of auditory stimulus processing at levels considered to be clearly beyond or above the limits imposed by the peripheral auditory system. Efforts to investigate such higher levels of processing of complex stimuli are nearly always based upon assumptions about perceptual and decision processes that limit the range of reasonably valid conclusions. Such assumptions are usually implicit and often not immediately recognized. To illustrate the critical role played by such implicit underlying assumptions, existing and new research on the perception of formant transitions in speech will be examined in terms of basic assumptions whose recognition can modify (and sometimes strengthen) conclusions about higher levels of perceptual processing. Discussion will focus on the implications of fundamental assumptions for the identification and demonstration of important principles of perceptual organizing (e.g., Gestalt, feature integration) and for testing hypotheses about alternative perceptual models, modes, or modules. [Research supported in part by NSF and AFOSR.]

Invited Paper presented at Acoustical Society of America,
Ottawa, Ontario, Canada
May 17, 1993 (2pMU3)

This paper is a little different than the preceding two papers in this session (Bregman, 1993; Darwin, 1993). Al Bregman provided a summary of the nature of modern auditory organization research as illustrated by some of the excellent, influential research conducted in his laboratory and by others. Chris Darwin then provided a summary of his creative research demonstrating the role of auditory organization principles in the perception of complex signals, including speech and music. I will shift focus to spend some time addressing some potential problems and pitfalls in modern research on auditory organization, and then suggest some possible solutions to the problems. I should note that in examining potential problems, one also can more fully appreciate the elegance of the research just described by Bregman and Darwin.

Much of the recent effort to evaluate the higher level processing of complex signals has been couched in terms of general perceptual principles derived by Gestalt researchers early in this century. This is a noble effort and the five sessions at this meeting speak to the modern importance of these efforts. However, there are some important limitations to this approach which must be kept in mind in attempting to draw strong conclusions.

Although working primarily with visual stimuli, Gestalt researchers faced several problems which can exist in modern auditory research efforts. One major problem was that the Gestalt research was based upon the analysis of static visual stimuli, and some of the Gestalt conclusions did not always readily generalize to dynamic stimuli. This problem is enhanced in attempts to apply Gestalt principles to audition, with the researcher often beginning with predictions (or even assumptions) based upon an analysis of a static visual representation of the auditory stimulus displayed in terms of time, frequency, and intensity. The visual representation of auditory stimuli makes the initial analysis potentially even more removed from perceptual reality than the early Gestalt work on visual perception. Thus, the predictions using general perceptual principles may sometimes be based upon artificial, static representations of the waveform which have not taken into account critical, perceptual limitations or important dynamic perceptual properties.

The second, and somewhat related problem with modern research applications, reflects a basic premise of the Gestalt approach. The Gestalt school developed as the reaction to previous research efforts which had made strong assumptions about the nature of the critical units of sensation and perception. One basic premise of the Gestalt approach is that the researcher needs to allow perceptual results to define the basic units of perception. This essential feature of the Gestalt approach is in contrast to the researcher making assumptions about those basic units and then proceeding as though those assumptions were valid. However, the Gestalt approach also can leave the researcher open to an inherent circularity where the identified principles are used on a post hoc basis to define the basic units used to demonstrate the principles. As a positive counter-example, I note that much of the work by Bregman, and the work referenced in Bregman's excellent book (Bregman, 1991) is careful about letting perception define the variables while avoiding the inherent problem of circularity.

What I hope to accomplish today is first to demonstrate (using speech research) some important examples of how assumptions about features or basic units of perception can lead to erroneous conclusions. I then will focus on some important new procedures which might be used to test notions about basic units of perception and thus avoid the problem of circularity.

In 1971 Mattingly and his colleagues published an important paper which has often been cited as a basis for contrast between the perception of speech (e.g., phonemes in CV context) and analogous nonspeech stimuli. I will spend a few minutes summarizing this study and its findings because, with the significant advantage of hindsight, that paper provides a basis to understanding what may, or may not, be perceptually analogous conditions, and allows us to begin to come to grips with adequate definitions of perceptual features or units.

The top portion of FIGURE 1 provides a summary of the stimuli used in the Mattingly paper. The stimuli were two formant synthetic CV syllables which varied in the onset frequency, and thus the nature of the second formant transition. Subjects were asked to label the stimuli and to perform an oddity discrimination task between pairs of stimuli differing in nominally equal step-size. The findings were fairly typical of the categorical perception literature. The labeling function (displayed as in the original paper

for one subject) exhibits relatively discrete labeling boundaries between the three categories of "b," "d," and "g." The discrimination function (as in the original study, pooled across subjects) exhibits peaks which roughly correspond to the location of the category boundaries.

The only part of the physical stimuli which varied was the second formant transition. Thus, analyzing the visual representation of the physical stimuli, it seemed obvious that any perceptual, nonspeech basis for the phonetic contrast must be carried by the transitions; all other parts of the stimuli were held constant. Subjects asked to label and to discriminate the isolated transitions failed to exhibit categorical perception; these findings were interpreted as supporting the notion that the identical stimulus component is perceived differently by speech and non-speech systems. Mattingly, et al. have made an implicit assumption that it is the isolated glide which is the most likely basic perceptual unit for the phonetic continuum. This assumption is intuitively appealing, but clearly ignores the basic premise of the Gestalt approach.

Notice that the Mattingly subjects might be perceiving a relative, rather than absolute, change in frequency across time. That is, perception may require the incorporation of the subsequent steady-state portion of the stimulus represented by the vowel formant. An ASA paper by Ralston and Sawusch (1983), and a more recent publication out of my laboratory, (Pastore, Li, & Layer, 1990) demonstrated that subjects experienced with sinewave stimuli yield relatively continuous perception for isolated transition, but categorical perception results when a following steady-state component is added. The three types of stimuli and the results for the short bleat stimuli are summarized in FIGURE 2. Notice that the patterns labeling and discrimination results closely parallel the syllable results reported by Mattingly. Therefore, it would appear that the critical perceptual component of stimuli varying in place of articulation may be a change in frequency leading into the relatively steady-state component. The notion of relative movement or change in formant frequency as being an important cue for place of articulation is something that Ken Stevens has been talking about for a number of years.

A number of early studies on glides or transitions built upon the procedures utilized in the pioneering study by Brady, Stevens, & House (1961), evaluating the pitch corresponding to isolated FM glides. The studies usually matching the pitch of a tone to the pitch of a previously presented FM glide. Typical results are that perceived pitch tends to be carried more by the offset frequency, with there being some differences between rising and falling glides. Notice that these studies parallel the original Mattingly, et al. assumption about the importance of isolated transitions (or chirps) as critical features, but have the additional assumption that the major acoustic role of a glide is in terms of an overall, static pitch quality.

A paper later in this session by Michael Hall (3pPU5) will present results of research with sinewave analogs. Pilot conditions (which Mike will not present) indicate that if one reverses the ordering of stimuli in a pitch matching study, matching a tone to a subsequent, rather than previous, glide, subjects will tend to match pitch based more on the onset, rather than offset frequencies in the glides. In other words, subjects seem to match pitch on the basis of the most temporally contiguous portions of the stimulus. Such a conclusion also implies that, when listening to glides, subjects are perceiving more than simply a single overall pitch quality, and may not even hear an overall pitch quality. The former conclusion is generally consistent with often ignored aspects of the early findings of Brady, House, and Stevens (1961). A later paper by Nabelek and Hirsh (1969) indicates that not only do subjects tend to perceive continuous changes in frequency, but they even impose perception of a glide when given a short time period between two steady-state stimuli differing somewhat in frequency. Therefore, perceptual results seem to argue that an important perceptual characteristic of glides is something more than, or different than, overall pitch quality.

Looking at the physical stimuli from Mattingly (FIGURE 1), notice that we appear to have good continuation between the glide and the following steady-state. Again, we need to be careful in not drawing conclusions about basic units of perception using only the static visual representation of the auditory stimulus. I don't want to steal too much from Mike Hall's later paper, but one of the questions is whether or not the presence of physical good continuation in the visual representation of the physical stimuli is equivalent to perceptual good continuation.

As shown in FIGURE 3, one can somewhat offset the ending frequency of the transition relative to the steady-state portion without disrupting the tendency to perceive the transition as appropriate for the phoneme normally defined in terms of continuity with the following portion of the stimulus. In Fig. 3 the rising (falling) transitions are perceived as equivalent even though the center frequency of one transition is equal to the starting frequency of the other. These results get back to the notion that an important feature for some consonants may be related to a perception of motion (or change) in frequency, with direction being important. However, we need to add a qualification, since Schouten (1986) found that the ability to perceive the direction of transitions may be quite limited, and that perceived direction need not correspond to physical direction of frequency change (reinforcing the point about letting perception define cues).

The notion of good continuation may still apply, but only when defined in perceptual, rather than physical, terms. Systematic research by Schouten (1985, 1986; Schouten & Pols, 1984), as well as a number of recent papers on perceptual limits for transitions may begin to provide some insights into when some aspects of transitions might begin to change in perceptual nature (Elliott et al., 1991; Dooley & Moore, 1988; von Wieringen & Pols, 1991). We suspect that, as Chris Darwin has demonstrated for similarity based upon fundamental frequency for vowel formants, a sufficient discrepancy in continuation can become perceptually salient, thus leading to segregation, rather than integration of perceptual elements. In fact, an earlier study by Repp and Bentin (1984) although interpreted somewhat differently, demonstrated that, with sufficient frequency offset, transitions do begin to perceptually segregate from the remainder of the synthetic CV syllables.

We now turn to the larger issue of defining critical aspects of perception--with the admonition that units of perception may be a function of level of perceptual analysis. One problem is a relative absence of research tools for determining possible basic units of perception, or basic perceptual features of complex auditory stimuli. This issue is the focus of the remainder of the talk.

In speech perception there is a phenomenon called normalization which has been explored by Pisoni and his colleagues. Normalization is really a modern analog of the classic notion of perceptual constancy. In speech it has been noted that perception of a CV syllable or a word persists despite changes in talker, speaking rate, and a number of other properties of the source event. Likewise, in music, one can perceive the equivalence of chords, or sequences of chords, played by different instruments. In the speech literature, normalization is argued to be an active process which takes time to implement. Irrelevant changes in speaker or instrument could be represented as added variability or noise, and thus lower signal-to-noise ratio, which should slow processing, but also reduce accuracy. In normalization, the perceptual system is assumed to be able to evaluate, and possibly even anticipate the nature of the irrelevant variability. The system then factors out the irrelevant variability, thus restoring accuracy, or imposing constancy, on perception. Notice that an anticipatory normalization process could potentially involve imagery. In fact, Crowder's excellent work on musical imagery is based upon the perceptual system retrieving some sort of an internal representation of an irrelevant stimulus properties which is to be factored out.

In order to effectively investigate normalization, the researcher really needs to begin with a reasonable conceptualization of the critical perceptual elements or features which are constant and those elements which are factored out. Alternatively, one might use normalization as a tool to begin to evaluate conjectures about essential perceptual features of auditory stimuli. This Friday, Jennifer Cho (SaMUS) will present a paper which looks at the relative roles of the attack and the upper partial in timbre normalization for natural and edited music stimuli. As part of this study, Jennifer evaluated the relationship between perceived similarity between stimuli defining instrument timbre and the normalization of instrument differences for the perception of chords. FIGURE 4 shows the reaction time measures for normalization was an inverse function of perceived similarity. Thus similarity scaling and normalization represent two very different procedures which can provide converging evidence in beginning to identify important perceptual features of music and speech.

Finally, I wish to turn to a model from cognitive psychology which was developed for visual stimuli, but which has potential implications for our understanding of auditory perceptual processing, and which is definitely related to the issue of the definition of features. In the early 1980s, Ann Treisman proposed her Feature Integration Theory (FIT) of perception. FIGURE 5 summarizes this theory. In the basic task a number of different stimuli are presented simultaneously to subjects. Stimuli exhibit different combinations of values along several dimensions such as size, shape, color, and location. According to FIT the values of features are preattentively extracted independently and in parallel, with rough tags (in a master map) for location. Therefore, search for single features among a stimulus array occurs in parallel and thus should be extremely rapid. For example, a yes-no task for the presence of a red or orange stimulus should occur very quickly and should be very little affected by the number of stimuli in the array (as long as the subjects are normal Trichromats and there is more than a few stimuli). Attention then is required for the serial task of *conjoining* the individual features to result in the perception of objects at different locations. However, the search for a conjunction of features, such as a red O, should be slower, especially as the number of stimuli is increased. Furthermore, when processing capacity is taxed, such as with a large number of stimuli, one should often find the perceptual system erroneously conjoining features presented at separate locations in the array. An illusory conjunction is the perception of an object whose features appear in the stimulus array, but never together.

Mike Hall, Wenyi Huang, Barbara Acker, and I have just completed a study (which we will report at a future meeting) using musical stimuli varying in instrument, pitch played, and the lateralized position in the head. Our results have the patterns of reaction time described by Treisman for conjoined feature searches. We also found illusory conjunctions of features at rates and levels of subject confidence which are equivalent to those reported by Treisman for visual stimuli.

Notice the FIT relies upon the appropriate definition of features, and that features are processed in a manner different than perceptual elements which are really conjoined features. In fact, some aspects of our results are best interpreted as our using complex, conjoined features, rather than basic perceptual features, for some of the "single-feature" pitch and instrument searches. Although it is interesting, and not really surprising, that one should find similar higher level processes for both visual and auditory (specifically music) stimuli, the techniques used to demonstrate FIT represent another approach to beginning to identify perceptual features.

I now return to my initial point that Gestalt researchers warned that proper conclusions about principles of perception need to begin with the identification of perceptual units defined by the perceptual system, and not simply by the researcher's intuition. When such units of perception have been defined properly, we will best understand those aspects of Gestalt principles which describe the perception of speech and music stimuli. The important point is that an adequate identification of the basic features of perception is critical for the development and evaluation of <u>all</u> models of complex auditory perception. In this paper we have looked at recent research developments which provide a number of new techniques which can be used to evaluate potential definitions of features and which can provide converging evidence in support of that type of evaluation.

Acknowledgement

Research described in this paper was supported in part by grants F496209310033 and F496209310327 from Air Force Office of Scientific Research and grant BNS8911456 from the National Science Foundation. THe opinions, results, and conclusions are those of the author and do not necessarily represent those of either granting agency.

References

Darwin, C.J., & Gardner, R.B. (1987). Perceptual separation of speech from concurrent sounds. In Schouten, M.E.H. (Ed.) <u>The Psychophysics of Speech Perception</u>. Nijhoff: Boston.

Dooley, G.J., & Moore, B.C.J. (1988). "Detection of linear frequency glides as a function of frequency and duration." Journal of the Acoustical Society of America,84, 2045-2057.

Elliott, L.L., Hammer, M.A., & Carell, T.D. (1991) "Discrimination of Second-formant-like frequency transitions," Perception & Psychophysics, 50, 1-6.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K., & Halwes, T. (1971). Discrimination in speech and nonspeech modes. Cognitive Psychology, 2, 131-157.

Nabelek, V., Nabelek, A.K., & Hirsh, I.J. (1970). Pitch of tone bursts of changing frequency. Journal of the Acoustical Society of America, 48, 536-553.

Pastore, R.E. (1981). Possible psychoacoustic factors in speech perception. In P.D. Eimas & J.L. Miller (Eds.), Perspectives in the Study of Speech. Hillsdale, NJ: Erlbaum. Chap 5.

Pastore, R.E., Li, X.-F., & Layer, J.K. (1990). Categorization of chirps and bleats; their similarity to speech. Perception & Psychophysics, 48, 151-156.

Porter, R.J., Cullen, J.K., Collins, M.J., & Jackson, D.F. (1991) "Discrimination of formant transition onset frequency: Psychoacoustic cues at short, moderate, and long durations," Journal of the Acoustical Society of America, 90, 1298-1308.

Repp, B.H., & Bentin, S. (1984). Parameters of spectral/temporal fusion in speech perception. Perception & Psychophysics, 36, 523-530.

Schouten, M. (1985). Identification and discrimination of sweep tones. Perception and Psychophysics, 37, 369-376.

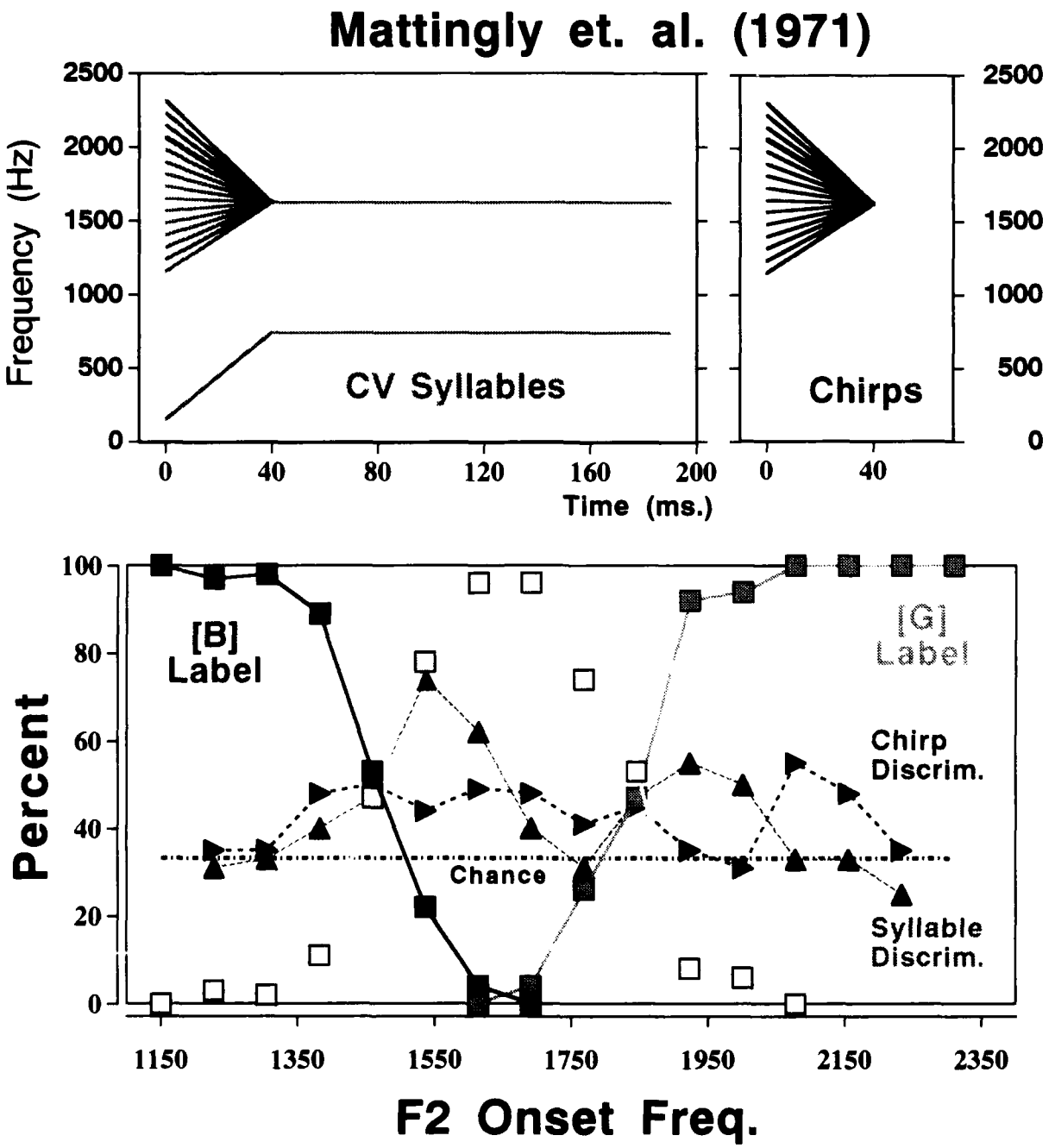
Schouten, M. (1986). Three-way identification of sweep tones. Perception and Psychophysics, 40, 359-361.

Schouten, M., & Pols, L. (1984). Identification and intervocalic plosive consonants: The importance of plosive bursts vs. vocalic transitions. In M. van den Broeck & A. Cohen (Eds.). Proceedings of the 10th International Congress of Phonetic Sciences, *Foris Publications, Dordrecht, 464-468.*

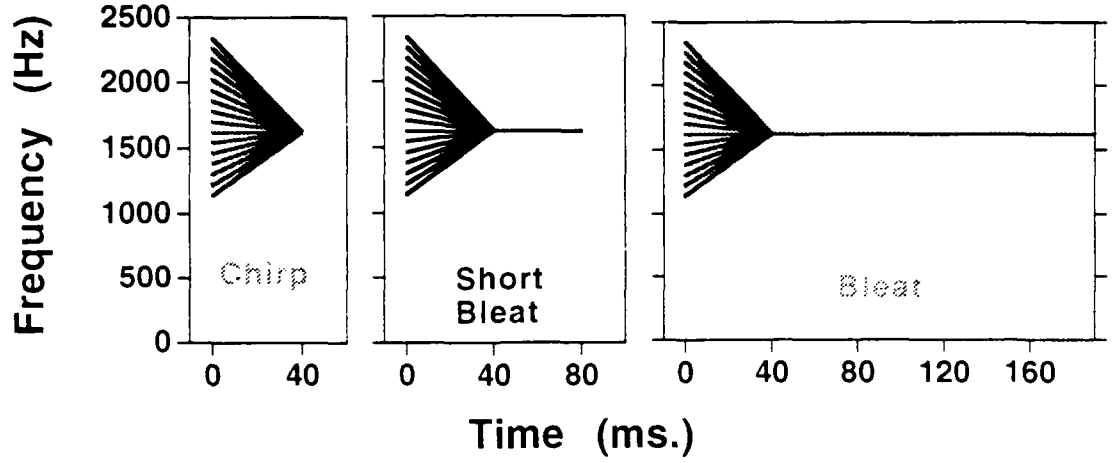
Treisman, A.M., and Gelade, G. (1980). A feature-integration theory of attention. Cognitive Psychology, 12, 97-136.

Wertheimer, M. (1958). Principles of perceptual organization. In D.C. Beardslee & M. Wertheimer (Eds.). Readings in Perception, 115-135. New York: van Nostrand.

Wieringen, A. van, & Pols, L.C.W. (1991). Transition rate as a cue in the perception of one-formant speech-like synthetic stimuli. Proceedings of the XIIth International Congress of Phonetic Sciences, Aix-en-Provence, Vol. 3, 446-449.


Mattingly et. al. (1971)
Frequency (Hz)
2500
2000
1500
1000
500
0
CV Syllables
Chirps
0
40
80
120
160
200
Time (ms.)
Percent
100
80
60
40
20
0
[B] Label
[G] Label
Chirp Discrim.
Chance
Syllable Discrim.
1150
1350
1550
1750
1950
2150
2350
F2 Onset Freq.

Pastore et al. (1990)



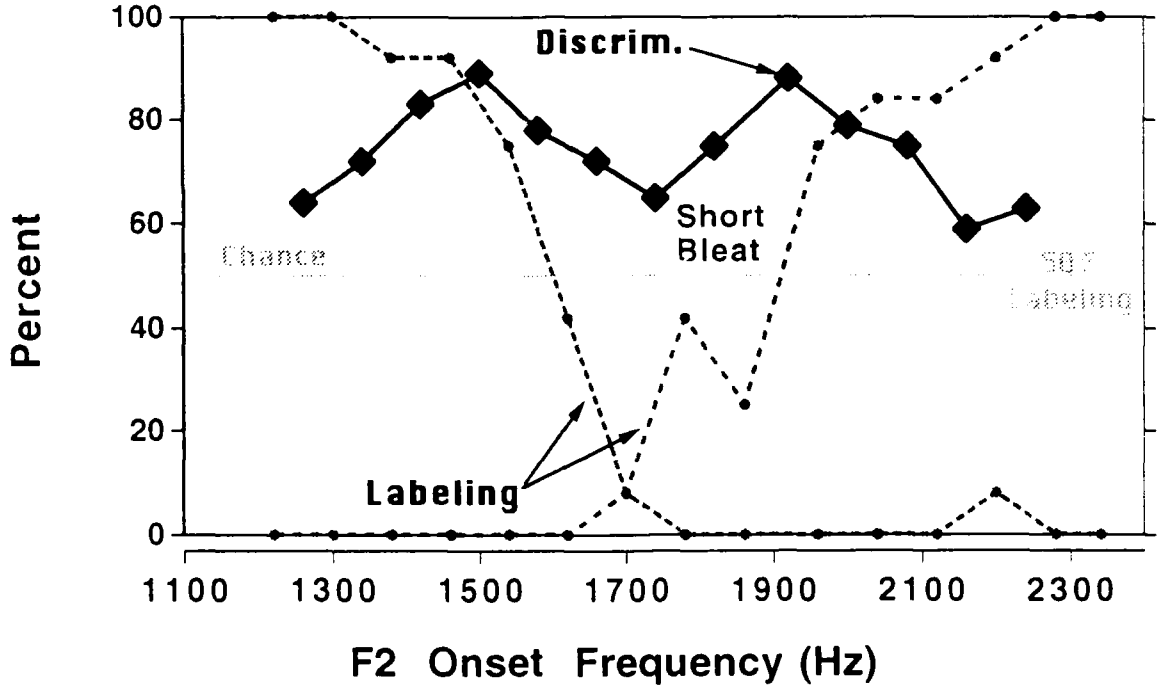


Figure 2

Figure 3

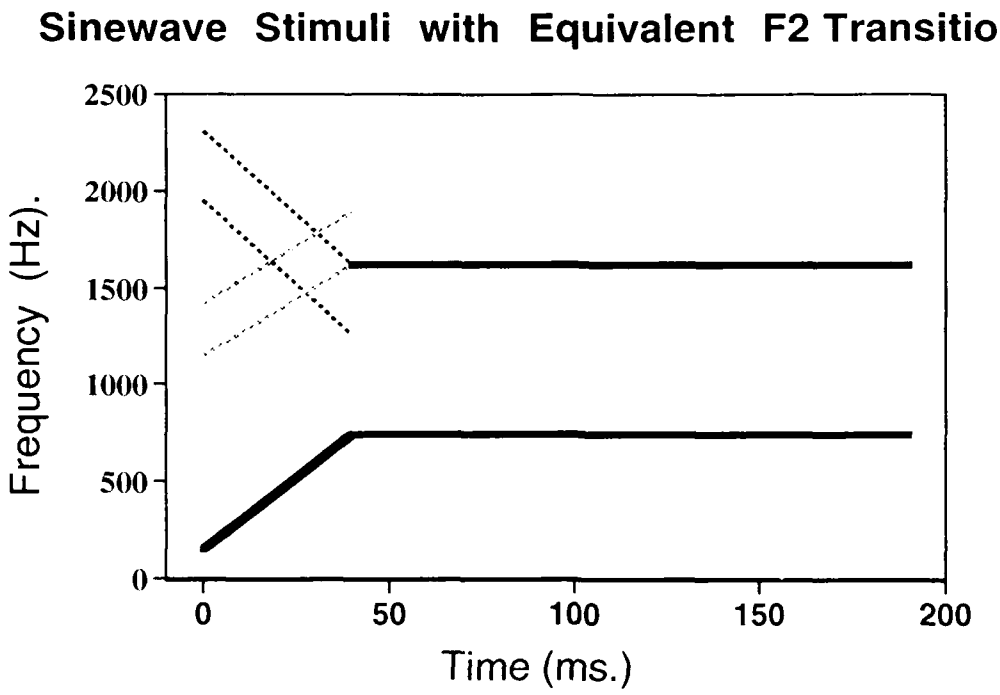
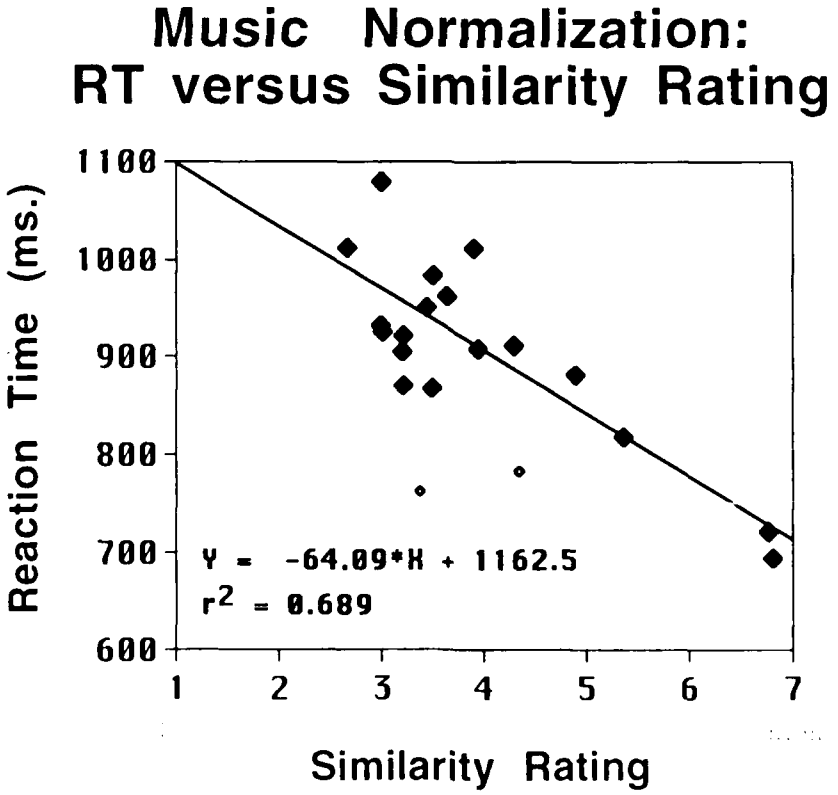
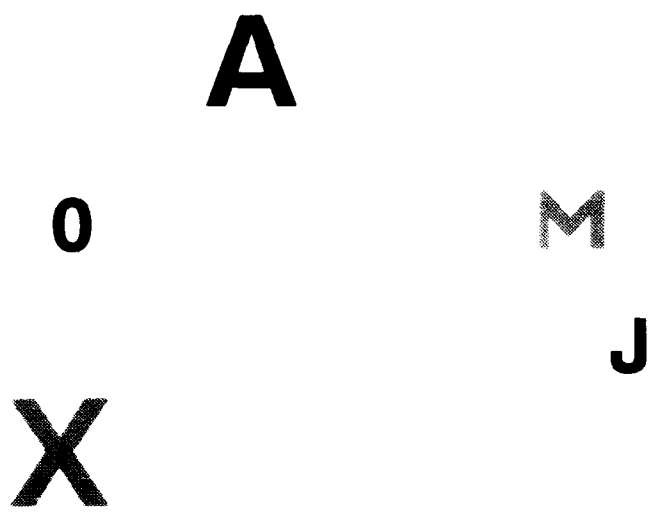


Figure 4





Feature Integration Theory

Single Feature Search:

- Red (valid) - Orange (invalid)
- X (valid) - Y (invalid)

Conjunction Search:

0 = Red "0" (valid)
Y = Orange "Y" (invalid)
J = Green "J" (illusory conj.)

Figure 5

Measuring the DL for Identification of Order of Onset for Complex Auditory Stimuli

Richard Pastore, Shannon Farrington, and Sajni Jassal

Abstract

Discussion of the approximately 20 ms. threshold for the identification of the order of onset of components of auditory stimuli has ranged from consideration of the absolute threshold (RL) as a possible factor contributing to the perception of voicing contrasts in speech to claims that the threshold is a methodological artifact. The current research investigates the identification of the temporal order of onset in terms of the Difference Limen (DL) for complex stimuli (modeled after CV syllables) which vary in degree of onset. The results provide clear evidence that the DL at relatively short onset differences (less than 25 ms) follows predictions based upon a perceptual threshold or limit. Furthermore, the DL seems to be a function of context coding of stimulus information, with both the DL and RL probably reflecting limits on the effective perception and coding of the short-term stimulus spectrum.

End of Abstract

Running Head: DL for Order Onset Identification

Somewhat over three decades ago Hirsh reported new findings on some temporal limits of perception which has run a full circle from being ignored to being widely cited, then often misunderstood, and finally again largely ignored. Hirsh (1959) found that there are a series or hierarchy of temporal limits on perception. For monaural (or diotic) presentation conditions, a difference of approximately 2 msec was required for detecting an onset ansychrony of two auditory stimuli (defining a threshold for simultaneity), but a difference of approximately 20 msec in onset was required to identify which of the stimuli had an earlier onset (defining a threshold for order of onset). This later Temporal Order Threshold (TOT) can be contrasted with a threshold of several hundred msec for correctly assigning the order labels to a sequence of four or more stimuli repeated in sequence (Warren, 1982). Moving in the other temporal direction, a difference of only a few microseconds is required to detect a difference in onset for the identical stimuli presented to the two ears, with this threshold really reflecting a difference in lateralization (Hirsh, 1974). Each of these thresholds has been replicated a number of different times by different researchers using a variety of psychophysical procedures. Hirsh's major point was that there are a number of different types of temporal limitations, with the two shorter thresholds probably reflecting sensory limitations, and the longer limits reflecting perceptual or even memory limitations on the processing of stimuli. These observations are important in their own right.

The focus of the current research is approximately a 20 msec threshold for TOT which, for a number of years, had been of extra interest because of its potential relationship to the perception of initial position stop consonants of English which are contrasted in voicing. Hirsh (1959) had observed that such stop consonants typically exhibit a labeling or categorization boundary when voicing onset is delayed by approximately 20 msec relative to the release or onset of the syllable. Hirsh (1959) conjectured that the TOT limitation may be an important underlying basis for voicing contrast. This conjecture makes a great deal of sense; the threshold implies that although subjects may be able to detect a difference in onset for various components of the complex signal (e.g., speech), an onset difference of at least 20 msec (in ideal laboratory situations) is required to reliably identify which component had an earlier (or later) onset. Thus, an initial consonant is voiceless if the higher frequency or asperated components are perceived as having an onset before either the lower frequency (F1) or the voiced components. Later research demonstrated categorical perception along a temporal onset continuum for noise buzz stimuli (Miller, Wier, Pastore, Kelly, and Dooling, 1976) and for pairs of tones (Pisoni, 1977). The pattern of categorical perception for the stimuli was fairly similar to that reported earlier for synthetic speech stimuli varying in Voice Onset Time (VOT). Thus, the 20 msec TOT was of interest both as one of a limited set of temporal limits on perception and because of its potential role in voicing contrast, the latter defining the TOT hypothesis.

Criticisms of the TOT Hypothesis

The first major criticism of the TOT hypothesis as the auditory basis for the perception of voicing contrast was based upon a systematic study by Summerfield (1982). Hirsh (1959) had found that TOT was independent of stimulus type. A number of later studies had demonstrated that the VOT boundary for American English syllables was a function of place of articulation, varying from 24-28 msec for labial stops to 28-52 msec (or more) for velar stops (for review, see Pastore, 1987a). Summerfield built upon these observations, demonstrating that TOT for tones and noise-buzz stimuli was relatively constant, and thus independent, of many of the frequency parameters which have been demonstrated to be important for changes in place of articulation. Furthermore, the TOT boundary was at a much shorter duration than reasonable.

If one adopts an extreme view about the TOT hypothesis, positing that voicing contrast is based solely upon whether or not the listener can perceive the order of onset of specific components, then the Summerfield study represents clear evidence against this hypothesis. However, a weaker version of the TOT hypothesis is still reasonable. According to the weaker version of the hypothesis, perception of an order of onset is an important contributing factor or cue for the perception of voiceless quality, but it is not the only cue. This multiple cue notion has a reasonable basis in the general findings that all speech categories are based upon the contribution of a number of different features or cues, with the perception of voicing contrast reflected in the action of up to 16 different cues (Lisker, et al., 1977). If voicing contrast, or any other phonetic contrast, were based upon any single factor or threshold, then the decades of research by outstanding scientists would have long ago succeeded in identifying the singular factor determining a phonetic contrast. Furthermore, there are logical grounds for conjecturing that temporal order limitation is a contributing factor to the perception of voicing contrast. Specifically, production with English stop consonant studies have found distributions of voiced stimuli which are skewed toward simultaneity, with very few tokens having VOT values even approaching the approximately 20 msec TOT, and with the distribution of voiceless tokens skewed toward higher VOT values, again avoiding the

Condition

ambiguous region near the TOT threshold limitation. One critical question, then, concerns what specific components of the stimuli might be perceived in an ordered fashion.

Although it is not necessary for voicing contrast to be based upon the ordered perception of the same limited set of stimulus components for all categories of place of articulation, a logical single starting point for a common temporal contrast would be the F1 cutback in which there is no activation of the low-frequency first formant resonance until after the onset of voicing, with the higher formants being activated within a very brief period following the release of the consonant. An alternative candidate might be whether or not the release burst is perceived as occurring clearly before the onset of voicing. The current research will focus on the first of these two conjectures.

Following the work of Summerfield (1982), several studies demonstrated that the TOT can be a function of stimulus parameters, especially when those stimulus parameters are consistent with stimulus properties exhibited in natural and synthetic speech stimuli. Hillenbrand (1984) demonstrated some variation in TOT for stimuli with dynamic changes in frequency at onset, thus mimicking with tonal stimuli the frequency changes associated with F1, F2, and F3 transitions. Pastore, et al. (1981; 1988) demonstrated significant changes in the TOT for stimuli with dynamic frequency transitions at onset coupled with variation in rise time and the presence of an initial release burst. For stimuli with slow rise times (typical of speech stimuli) and strong initial release bursts, TOT, measured using a two alternative force choice procedure, was found to be as long as 40 msec for certain types of stimuli, especially those patterned after velar stimuli contrasted in voicing. These results would all seem to support the weak version of the TOT hypothesis.

Recent Criticisms of TOT

In the late 1980s there were two new criticisms of the TOT hypothesis, both of which argued against the existence of an approximate 20 msec threshold for temporal order identification. The argument by Rosen and Howell (1987) is a theoretical one based upon methodological issues. The argument by Kewley-Port, Watson, and Foyle (1988) is based upon empirical work under minimal uncertainty conditions. In addition, the excellent text on hearing by Moore (1989) simply does not recognize a limitation on temporal order, instead citing only the 2 msec boundary for simultaneity and the 200 msec boundary for labeling the individual components in a sequence of stimuli. These arguments about the absence of a 20 msec temporal order threshold ignore the replication of the threshold using a number of different psychophysical procedures. However, the heart of each of the major criticisms will be addressed before summarizing the current research effort.

Hirsh Methodology

Rosen and Howell (1987) argue that the Hirsh (1959) procedure is really a labeling task, and that when the original Hirsh results are replotted in terms of labeling, one obtains a 50% category boundary at approximately onset synchrony (0 msec onset difference), with no indication of any threshold or limitation at approximately 20 msec.

Hirsh used standard psychophysical methodology from the time to measure a difference threshold (Difference Limen or DL). Understanding the assumptions underlying the measurement of a DL, and the limitations with which Hirsh (1959) had to deal, one sees that the Rosen and Howell argument is based upon the implicit and testable assumption that the relevant DL does not exist. Relevant notions about measuring the DL can be found in the important text from the time of Hirsh's research (e.g., Osgood, 1953; Woodworth and Schlosberg, 1954) and are summarized in Figure 1 (panel A). According to classic notions, there is a region around the Point of Objective (or Physical) Equality (POE) within which subjects perceive an equivalence to the standard or cannot reliably identify a difference from the standard. This region of perceptual equality is called the interval of uncertainty, its midpoint is called the Point of Subjective Equality (PSE), and the DL is half of the size of the interval of uncertainty. In the case of the research by Hirsh (1959), there really are two regions of uncertainty surrounding temporal onset synchrony. There is a very narrow region for the perception of synchrony (this interval of uncertainty equals 4 msec) and a broader region within which subjects cannot reliably identify the order of onset (this interval of uncertainty equals 40 msec). Beyond this region in each direction subjects can reliably discriminate stimulus differences. As with any threshold, there is a statistical distribution of response probabilities, rather than a quantum or discrete change in the probability of response.

One method for measuring the DL is to ask subjects to adjust the stimulus magnitude either from a state of clear perceptual difference to an initial state of equality, or by beginning with the stimulus of physical equality and ask the subjects to adjust the stimuli until they can first perceive the designated perceptual difference (e.g., order of stimulus onset). Either method of adjustment would define the two limits of the interval of uncertainty which define the boundaries between three response regions (perceptually lower, perceptually equal, and perceptually higher). Unlike thresholds defined in terms of intensity, a TOT requires a discrete trial situation where it is difficult to manipulate onset difference, especially using the technology available in the 1950s. Hirsh thus was forced to use the method of constant stimuli. For methods of constant stimuli the three response regions can be mapped using either a 3-category or a 2-category method (Osgood, 1953).

Insert Figure 1 about here

Fig. 1A shows typical results for measuring the DL using a 3-category version of the method of constant stimuli. The abscissa represents an arbitrary designation of the stimulus continuum, with zero being the stimulus standard (POE). In the case of a temporal order continuum, zero would represent onset synchrony, with the continuum representing (in totally arbitrary units) the relative onset of the two components to the stimulus (e.g., lower frequency having an earlier, versus later, onset). The three curves, each representing an assumed underlying category, are plotting the relative probability for responding to category A or B

(discriminable differences) and the uncertain (or equality) category. In this hypothetical illustration the three response probabilities are actually based upon normal (uncertain category) and cumulative Normal or Gamma (categories A and B) distributions, with the relative proabilities scaled to sum to unity for each stimulus. Using classic psychophysical methodology, the interval of uncertainty can be estimated in terms of the stimulus values yielding 50% identification for categories A and B, with the DL being half of this interval. Since the illustration in Figure 1a has plotted an ideal interval of uncertainty which is symmetric around the POE (i.e., PSE = POE), the value of the stimulus at the upper limit represents the DL, and is so indicated in Fig. IA.

However, *in measuring TOT the separate threshold for simultaneity represents a major problem* in defining allowable responses for subjects. On a conceptual basis, as one increases the onset difference (in either direction) from onset synchrony, perception changes from (1) synchrony to (2) asynchrony coupled with an inability to identify order of onset, and finally to (3) a clear ordered onset. The simplest and most understandable instructions to the subjects is to indicate the order of onset, measuring threshold based upon correctness of response. This approach, which was used by Hirsh (1959), follows the standard 2-category method for measuring the DL (Osgood, 1953). In the case of Hirsh's Temporal Order study, the targeted middle category is not one of simultaneity (4 ms wide), but rather the broader (40 ms) range for inability to correctly perceive order of onset; the middle response category cannot indicate equality of onset, but rather a failure to accurately identify order of onset. Thus, the 3-category procedure therefore could not be effectively implemented.

Insert Figure 1 about here

The logical, alternative procedure (e.g., Osgood, 1953) is to use only two-response categories (the "low" and "high" categories from Fig. 1a). The resulting psychometric functions, derived from the three distributions found in Fig. 1a, are plotted in the lower panel (b) of Fig. 1). The "low" responses should occur either when the subject perceives a "low" event (reflected in the p (low) curve in Fig. 1a) or when the subject jointly fails to perceive either category (reflected by the uncertain or neither category in Fig. 1a) and guesses "low." Assuming an equal probability of guessing sampling from the uncertain or neither category, the probability of responding "low" for any given stimulus therefore can be derived from Fig. 1a and should be equal to the p (low) plus one-half of p (neither). The resulting 2-category psychometric functions have been plotted in Fig. 1b. Notice that the two psychometric functions no longer resemble a cummulative Normal distribution, but rather have distinctive deviations (e.g., are shallower, rather than steep, in slope) for stimuli falling within (and near) the interval of uncertainty. The 50% point for these 2-category psychometric functions represents the point of subjective equality (PSE) which, (by definition) should at least approximate the POE, and thus onset synchrony.

Because the 2-category procedure forces guessing, the points which had been represented by the 50% performance values in Fig. 1A now must be estimated from the 75% points in Fig. 1B (e.g., 50% correct responding plus 50% guessing for the remaining trials). The 75% criterion for the 2-category procedure yields the same DL estimate as the 50% criterion with the 3-category procedure (as illustrated in Fig. 1). Rosen and Howell (1987) have made an implicit assumption that the middle category (interval of uncertainty) does not exist and have taken as supporting evidence the correspondence between the PSE and POE.

The functions in Fig. 1B also provide us with an important insight into the category structure being used by subjects. The deviation from a reasonable cummulative Normal distribution found in Fig. 1b can be interpreted as a clear indication that the subjects have an additional, intervening category which they are distributing between the two allowable response categories. Hirsh (1959) used the 2-category procedure and his results (scaled in terms of Z-scores) to exhibit this critical characteristic (a linear function with a bend or elbow). The Hirsch findings thus indicate the existence of the implied intervening category. [It also should also be noted that it is fairly typical to find such deviations from a reasonable cummulative Normal distribution in the labeling results for a number of different speech continua.

Sensory or Perceptual Limit

Kewley-Port, Watson, and Foyle (1988) have argued from a different perspective that there is no temporal order threshold, at least not at the limits of sensory capabilities. Their study reports on a lack of evidence for any temporal order limit of approximately 15 to 20 msec under minimal uncertainty conditions using highly practiced subjects. The Kewley-Port, et al. study is a solid research project on perception under minimal uncertainty conditions. Although we have questioned whether there is truly a lack of evidence in their results for a threshold at approximately 15 to 20 msec (Pastore, 1988; see also Watson & Kewley-Port, 1988), the important issue for the current study is that Hirsh (1959), and other subsequent research on temporal order perception (e.g., Miller, et al., 1976), have clearly argued that the temporal order limitation is perceptual, and does not represent a sensory limitation (we will return later to the issue of the nature of such a perceptual limitation). Therefore, the issue of whether or not one finds evidence for a temporal order identification threshold of approximately 15 to 20 msec at the limits of sensory capabilities, although interesting in its own right, is really irrelevant to perceptual research focusing on this threshold, and its possible role in the perception of voicing contrasts. Furthermore, when a finding has been often replicated under a wide variety of more natural conditions, a single failure to replicate under extreme conditions cannot falsify the more typical findings.

Kewley-Port, et al. also addressed temporal order perception (under minimal uncertainty conditions) in terms of the size of the Weber fraction; this represents an important approach to studying TOT. If there is no TOT then the function relating t to t should reflect a Weber function, and thus shall be linear with a positive slope corresponding to the Weber constant. From a separate, theoretical basis, Pastore (1987b) had also addressed the nature of TOT as an absolute perceptual threshold, incorporating the notion of a Weber fraction. This analysis conjectures a two stage function relating t to t. For relatively small (below threshold) onset time differences, the size of a just noticeable difference in temporal order of onset, t, should equal the difference between

the given stimulus, t_i, and the absolute threshold for order discrimination, t_b (e.g., any subliminal stimulus should be just discriminable from the stimulus which first exceeds threshold). This relationship is described by the formula

$$t = (t_b - t_i) + c \quad (\text{for } t_i < t_b) \tag{1}$$

There really are two different values of t_b. One value of t_b represents the threshold for simultaneity and is approximately 2 msec. The other value is the temporal order identification threshold which, for relatively stationary stimuli such as those used by Hirsh (1959), Pisoni (1977), and Miller, et al. (1976) would be approximately 20 msec. However, with dynamic stimuli similar to those used by Pastore, et al. (1988), t_b would be at a larger temporal order difference, possibly up to 45 msec. In this equation the constant, c, represents noise or variability in the system. This formula based upon quantal notions of absolute threshold, predicts that the size of the just noticeable change in temporal onset should decrease linearly (slope of -1.0) for temporal onsets up to t_b, the threshold for temporal order identification. Because we are dealing with a psychophysical procedure based upon probability distributions, we ran a simulation of observer behavior in a 2IFC task measuring t. The simulation assumed that judgement is based upon whether or not threshold (t_b) is exceeded (as in Eq. 1). In this simulation underlying Gaussian distribution of noise was assumed, with the stimuli spaced at equal Z-score distance. The additional constraint on the simulation was that subjects would guess (p = 0.5) when stimuli were both either below or above threshold (there is the added implicit assumption that two compared supra-liminal stimuli differs by less than the DL for supra-liminal stimuli]. Psychometric functions were generated, then used to estimate the 75% estimate of t for each value of t_i up to a value of t_b. The simulation indicated that the measurement procedure should result in a linear function, but with a slope of approximately -0.5 and with t = 0 at a value of t_i which exceeds by a small amount theoretical value of t_b.

This derivation provides one contrast in predictions: whether the function relating t to t_i (for smaller values of t_i) has a slope of -0.5 or less (threshold model) or has a positive slope (Weber function). If we assume the validity of Weber's law for temporal differences above this threshold, then the remainder of the differential sensitivity function should follow the equation:

$$t = k(t_i - t_b) + c \quad (\text{for } t_i > t_b) \tag{2}$$

In Eq. 2, k is the Weber constant, with all of the other parameters being equivalent to those in Eq. 1. Both equations for the threshold notion predict that the difference threshold should reach a minimum value of c when the standard, t_i, approximates t_b. Eq. 1 predicts a linear decrease in the size of the just noticeable difference up to t_b, with Eq. 2 predicting a linear increase in the size of the DL for values of $t_i > t_b$. Thus, the second major difference in predictions is that the threshold model requires two linear functions with an intersection representing a minimum value of t near threshold (t_b), whereas the no threshold model based upon a Weber fraction predicts a single linear function with a positive slope.

There are a number of different issues to be addressed in the current research. The first issue is whether discrimination results for the perception of order of onset follow the pattern predicted by the functioning of an absolute threshold. The second issue relates to the degree to which context coding plays a role in the identification of order of onset. This issue builds upon notions which were originally developed by Durlach and Braida (1969) and have been more recently applied to speech in the work of Macmillan, Braida, & Goldberg (1987; Uchanski, Millier, Reed, and Braida, 1992). This second issue is of interest not only in its own right, but also because recent work by Schouten and van Hessen (1992) reported results for the discrimination of phonemes contrasted in place of articulation which seem to indicate a lack of trace coding (thus, use of only context coding), with a lack of differences predicted on the basis of the psychophysical task employed.

General Methods

Subjects

A total of eight subjects were hired for this task. All had normal hearing and were native speakers of English. All were student age and considered participation to be a part-time summer position. Subjects were run in commercial sound chambers and listened to the binaural stimuli over TDH-49 earphones.

Stimuli

Two sets of stimuli were synthesized (10 kHz sample rate, 12-bit converter). The stimuli were sinewaves, with both frequency and amplitude changing as a function of time in a manner which was consistent with that described for the CV syllables used by Volitas and Miller (1992). All of the stimuli had two major components which corresponded to the F1 and F2 resonances. The F1 component began at 180 Hz and rose linearly to a steady state of 330 Hz over the first 20 ms of the tonal portion of the stimulus. The frequency of this low frequency (F1) component remained at 330 Hz for 180 ms, then declined to 300 Hz over the last 100 ms. The frequency of the higher frequency (F2) component followed a similar temporal pattern of frequency change. It either rose from an initial frequency of 1800 Hz to a steady-state frequency of 2200 Hz, or fell from an initial frequency of 2400 Hz to the same steady-state frequency. Therefore, the two types of stimuli (Rising and Falling F2 stimuli) differed in terms of the onset frequency of the higher frequency component. A given subject only listened to the two rising or the falling type of stimuli (Rising or Falling F2), and ran under both fixed and roving discrimination conditions. Therefore, we used a within subject design for stimulus type.

The amplitude of the original F1 stimulus and the F2 stimuli rose as a linear function of duration over the first 45 ms After maintaining a steady-state amplitude for 155 ms, the amplitude fell to zero (ground) over the last 100 ms. A basic onset

time continuum was created by replacing the first n ms (n = 5, 10, ..., 120) of the F1 stimulus with silence, then having the amplitude grow linearly over the next 4 ms to the amplitude of the original stimulus at that point in time. The onset-time continua varied in steps of 5 ms, with 0 ms (onset synchrony) defined in terms of the original F1 and F2 stimuli. The stimuli then were modified to produce nonspeech stimuli which better resembled CV sullables. An initial 5 ms noise burst plus 10 ms of silence then was added to the beginning of each F2 stimulus. The noise burst was a segment of white noise band-pass filtered at two-thirds of an octave centered at either 1800 Hz (Rising F2) or 3000 Hz (Falling F2), based upon an attempt to provide an analog to a stop consonant release burst. The F1 component had a 15 ms segment of silence (ground) added to its beginning. The given F1 and F2 components were produced simultaneously by separate DAC channel (sharing a common time base), then mixed as analog signals to create stimuli which differed in the onset of the (delayed) F1 component relative to the F2 component. A nominal onset difference of n ms thus consisted of a 5 ms burst of noise 10 ms of silence, n ms of the F2 component only, and then the F1 and F2 combined for (300-n) ms. If timing were specified in terms of VOT as measured between the release of the syllable and the onset of voicing (marked by the end of the F1-cutback), an n ms offset difference for the current stimuli would correspond to a VOT of (n + 15) ms.

Procedure

Both experiments used a 2IFC task. For any given trial two stimuli were presented, each with a different onset time (e.g., 5 ms vs. 45 ms). The computer randomly determined which of the two intervals contained the stimulus with the longer onset difference, as well as randomly selecting the specific longer stimulus from a pre-determined set. The task of the subject was to indicate, by button-press, which of the two intervals contained the stimulus with the longer onset difference. In the minimum uncertainty condition, the stimulus with the shorter onset time was constant throughout the block of trials, thus allowing generation of a psychometric function for only that base stimulus onset time. In the roving condition the stimulus with the shorter onset time differed from trial-to-trial, with the other stimulus always being longer in onset time.

All subjects were trained initially with the extreme values (5 and 120 ms) of the given stimulus type assigned to them. On each trial one of the stimulus had an onset difference of 5 ms and the other had an onset difference of 120 ms, with stimulus order randomly determined. The subjects were given feedback. They ran in short blocks of trials until they could perform the task with a high degree of accuracy (at least 90 percent). The individual subjects then were run with the 5 ms standard in a fixed condition with comparison stimuli differing in large steps across the onset continuum. The goal of this next condition, which also included feedback, was to evaluate the approximate location for the steep portion of the psychometric function (e.g., determining the approximate size of the difference limen).

Experiment 1: ISI

The first experiment evaluated the effects of varying ISI in the discrimination of temporal order onset. If the distinction between trace and context coding is important for the discrimination of temporal order of onset, then one would expect both trace and context coding for short inter-stimulus intervals and only context coding for longer inter-stimulus intervals.

Methods

Subjects

This experiment began with four subjects to each of the two stimulus types. As the experiment progressed, some subjects left the experiment to take other positions, and were not replaced. Two subjects resigned after the orientation conditions, leaving six subjects to complete Experiment 1. One of these subjects then resigned leaving five subjects to complete Experiment 2.

Procedure

A fixed discrimination procedure was used to generate a psychometric function for each of four time intervals (ISI) between the pair of stimuli on each trial. On each trial the minimum (standard) onset difference was 5 ms, with longer onset differences for the comparison stimuli. Linear regression of Z-score transformed data were used to estimate the 75% difference limen. This condition was repeated with a different ISI until difference limen were estimated for ISI values of 100, 300, 500, and 1500 ms. Order of running was counter-balanced across subjects, except that all subjects initially ran with a 500 ms ISI.

Insert Figure 2 about here

Results and Discussion

The results are summarized in Figure 2, which plots the size of the difference limen as a function of ISI. It is clear that the DL for the Falling F2 stimuli (filled symbols) is much smaller than for the Rising F2 stimuli. Although for four of the six subjects the DL is larger at the smallest (100 ms) values of ISI than at somewhat larger values of ISI, the small average difference is definitely not significant. Furthermore, the results would tend to indicate that if trace information is available at 100 ms, it serves to hinder, rather than help, the subjects. The psychophysical function is essentially flat. Therefore, it would appear that subjects are not changing their strategy in terms of the information available and employed as a function of ISI. These results would seem to indicate that the distinction between trace and context coding is not important in the identification of temporal order of onset. This conclusion will be discussed after the next experiment.

Experiment 2: Testing Models

The second experiment provides a direct test of whether there is a perceptual threshold by examining the discrimination function in terms of a two-stage versus a singular linear relationship. In addition, this experiment compared the size of the difference limen as a function of fixed versus roving procedure, thus further evaluating performance of this task in terms of the underlying coding of information.

Methods

Subject

A total of five subjects completed this experiment. All had participated in the early experiment. Three of the subjects ran the conditions with the Falling F2 stimulus, with two of the subjects running with the Rising F2 stimulus.

Procedure

In this experiment ISI was fixed at 500 ms. Psychometric functions were generated for each subject under both fixed and roving discrimination conditions for base stimulus differences from 5 to 85 ms in 10 ms steps, except that the fixed discrimination task at 75 ms was not run. The difference limen was estimated from the 75% point of the linear regression line of the Z-transformed psychometric function.

Results and Discussion

The roving and fixed discrimination results are summarized in two panels of Figures 3 and 4. It is clear that the relationship between the difference threshold and the onset difference is best described by a two-stage linear function. For relatively short onset differences, the DL is a linear decreasing function of onset difference. The average slope of this segment of the psychometric function is -0.55, and thus is consistent with the predictions of the threshold model. The value of the threshold appears to be at approximately 25.6 ms for the Falling F2 stimulus and approximately 45 ms for the Rising F2 stimulus. Converting these values of the DL to comparable values of VOT (adding burst and silence duration) yields an average phoneme equivalent of approximately 41 ms. Furthermore, when compared across the full onset time continuum and across the fixed and roving tasks, the size of the DL for the Falling (Figure 3) and Rising (Figure 4) F2 stimuli are equivalent.

At longer onset differences the size of the difference limen grows very slowly as a function of onset difference, with the function having a slope of only slightly greater than 0 (mean = 0.05). Therefore, average discrimination relative to the intial difference in onset is approximately constant. Thus, detection of differences in onset for stimuli above the threshold for detecting onset asynchrony does not exhibit a Weber function, but instead grows very slowly as a function of onset difference. The lack of a Weber relationship for the longer onset differences is surprising. However, recently Zera and Green (1993) reported results on the detection of differences in temporal onset for complex stimuli. Although Zera and Green were investigating a different type of task, the results do parallel those reported here. We believe that the recognition of order of onset is taking place at a different higher level of perceptual processing, it is not surprising that the shape of the two functions could be similar. Finally, the results for the fixed discrimination task are not significantly better than the roving discrimination, at least when measured in terms of the size of the difference limen.

Insert Figures 3 and 4 about here

General Discussion

The subjective report of our subjects are quite consistent with our own impressions about the temporal order task, and with the measurement of other differences based upon temporal properties of stimuli. The subjects do not seem to be responding directly to the temporal properties of the stimuli, but rather to perceptual changes in the stimuli which are correlated with, or a function of the temporal properties. Examining our stimuli from this perspective provides some insight into the nature of the temporal order task. A brief stimulus which is less than 10 ms in duration is heard as a click. If one increases the duration of the stimulus beyond approximately 10 ms, the stimulus begins to acquire a pitch-like quality, with pitch achieved only for longer stimuli. This perceptual phenomenon is related to the spectrum of brief signals where the effective signal band width is an inverse function of duration (the sinc function). This well-known observation can be generalized to the task of identifying which of two stimuli had an earlier onset. With onset differences of less than a specific value, the earliest portion of the initial stimulus provides the subject with only broad-band information which is insufficient to identify the stimuli. Stated another way, subjects require more than 10 to 20 ms of a signal in order to begin to perceive any pitch-like quality, and thus the temporal onset difference must exceed this value for the subject to begin to have sufficient information to identify which of the stimulus components had an earlier onset. If the stimuli have slow rise times or are dynamically changing in frequency composition, it is not surprising that longer onset differences would be required for subjects to identify which of two stimuli had an earlier onset.

When one uses highly practiced subjects with a limited set of conditions, one would expect to be able to push the threshold toward much shorter onset differences, and we have reported such results in the past (Pastore, Harris, Kaplan, 1981). Conversely, from dealing with stimuli which vary in composition (e.g., natural speech), it should not be surprising the limits on temporal order identification would be at a much longer onset asynchrony.

In the very real sense, threshold for temporal order identification could be considered a perceptual, or even cognitive, limit on performance. In the temporal order identification task a subject must acquire sufficient information about the stimulus with the earlier onset to be able to consistently apply the appropriate label to the stimulus. Although the threshold is specified in terms of temporal duration, the real goal for the subject is the acquisition of sufficient stimulus specification to perform the identification task. With the longer values of TOT, the limits on perception of pitch is no longer relevant. Instead, subjects probably are required to make a judgement based upon the temporal duration of the initial stimulus component. For these longer stimuli, the temporal order task becomes equivalent to the temporal discrimination task studied by Zera and Green (1993).

Using this conceptualization, we can now turn to the issue of context versus trace coding. It was highly doubtful that subjects can store trace representation of the onset of the initial stimulus for a sufficient duration so that it may be compared directly with the onset of the second stimulus. In addition to masking or interference from the later portion of the first stimulus, the task of the subject for short onset differences would be equivalent to attempting to compare the trace of two clicks or tone pips. Once the duration of the onset difference is sufficient, subjects can apply a label to the stimulus in terms of the onset difference, and then can perform the discrimination based upon whether or not the second stimulus falls in the same perceptual category. Therefore, subjects really are using only context coding, and are not using trace coding in performing the discrimination task at least for onset differences up to threshold.

Bibliography

Blumstein, S.E., and Stevens, K.N. (1981). The search for invariant acoustic correlates of phonetic features. In P.D. Eimas and J.L. Miller (Eds.), Perspectives in the Study of Speech. Hillsdale, N.J.: Erlbaum.

Hillenbrand, J. (1984). Perception of sine-wave analogs of voice onset time stimuli. Journal of the Acoustical Society of America, 75, 231-240.

Hirsh, I. J. (1974). Temporal order and auditory perception. In H. R. Moskowitz, B. Scharf & J. C. Stevens (Eds.), Sensation and measurement (pp. 251-258). Dordrecht, Holland: Reidel.

Hirsh, I.J. (1967). Information processing in input channels for speech and language: The significance of serial order of stimuli In Brain Mechanisms Underlying Speech and Language. (NY: Gruene & Stratton). 21-39.

Hirsh, I.J., and Fraisse, P. (1964). Simultanéité et succession de stimuli hétérogènes. L'année Psychologique, 64, 1-19.

Hirsh, I.J. (1959). Auditory perception of temporal order. Journal of the Acoustical Society of America 31, 759-767.

Kewley-Port, D., Watson, C.S., and Foyle, D.C. (1988). Auditory temporal acuity in relation to category boundaries: Speech and nonspeech stimuli. Journal of the Acoustical Society of America, 83, 1133-1145.

Li, X-F., and Pastore, R.E. (1992). Evaluation of prototypes and exemplars for a phoneme place continuum. In M.E.H. Schouten (Ed), Audition, Speech and Language, Berlin: Mouton-De Gruyter, 303-308.

Lisker, L., Liberman, A.M., Erickson, D.M., Dechovitz, D., and Mandler, R. (1977). On pushing the voice-onset-time (VOT) boundary about. Language and Speech, 20, 209-216.

Macmillan, N.A., Braida, L.D., and Goldberg, R.F. (1987). Central and peripheral processing in the perception of speech and nonspeech sounds. In Schouten, M.E.H. (Ed.) The Psychophysics of Speech Perception. Nijhoff: Boston, 28-45.

Miller, J.D., Wier, C.C., Pastore, R.E., Kelly, W.J., and Dooling, R.J. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. Journal of the Acoustical Society of America, 60, 410-417.

Moore, B.C.J. (1989). An Introduction to the Psychology of Hearing. (NY: Academic Press).

Osgood, C.E. (1953). Method and Theory in Experimental Psychology. (NY: Oxford).

Pastore, R.E. (1988). Burying straw men without graves: A reply to Kewley-Port, Watson, and Foyle (1988). Journal of the Acoustical Society of America, 84, 2262-2266.

Pastore, R.E., Layer, J.K., Morris, C.B., and Logan, R.J. (1988). Temporal order identification for tone/noise stimuli with onset transitions. Perception & Psychophysics, 44, 257-271.

Pastore, R.E. (1987a). Possible acoustic bases for the perception of voicing contrasts, In M.E.H. Schouten (Ed.). Psychophysics of Speech Perception, (Boston: Martinus Nijhoff), 188-198.

Pastore, R.E. (1987b). Categorical perception: Some psychophysical models. In S. Harnad (Ed.). Categorical Perception. (New York: Cambridge University Press). Chap. 1.

Pastore, R.E., Harris, L.B., & Layer, J.K. (1981). Temporal order identification: Some parameter dependencies. Journal of the Acoustical Society of America, 71, 430-436.

Pastore, R.E. (1981). Possible psychoacoustic factors in speech perception, In P.D. Eimas and J.L. Miller, (Eds.). Perspectives in the Study of Speech, (Erlbaum, Hillsdale, N.J.), Chap 5.

Pastore, R.E., Harris, L.B., and Kaplan, J.K. (1981). Temporal order identification: Some parameter dependencies. Journal of the Acoustical Society of America, 71, 430-436

Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two-component tones: Implications for voicing perception in stops. Journal of the Acoustical Society of America, 61, 1352-1361.

Rosen, S., and Howell, P. (1987). Is there a natural sensitivity at 20 ms in relative tone-onset-time continua? A reanalysis of Hirsh's (1959) data. In M.H.E. Schouten (Ed.), Psychophysics of Speech Perception. Boston: Nijhoff. 199-209.

Sawusch, J.R. (1992). Auditory metrics for phonetic recognition. In M.E.H. Schouten The Auditory Processing of Speech: From Sounds to Words. NY: Mouton de Gruyter. 315-321.

Schouten, M.E.H., & van Hessen, A.J. (1992). Different discrimination strategies for vowels and consonants. In M.E.H Schouten The Auditory Processing of Speech: From Sounds to Words. NY: Mouton de Gruyter, 309-314.

Soli, S. (1983). The role of spectral cues in discrimination of voice onset time differences. Journal of the Acoustical Society of America, 73, 2150-2165.

Summerfield, Q. (1982). Differences between spectral dependencies in auditory and phonetic temporal processing: Relevance to the perception of voicing in initial stops. Journal of the Acoustical Society of America, 72, 51-61.

Uchanski, R.M., Millier, K.M., Reed, C.M., & Braida, L.D. (1992). Effects of token variability on resolution for Vowel Sounds. In M.E.H. Schouten The Auditory Processing of Speech: From Sounds to Words. NY: Mouton de Gruyter, 291-302.

Volatis, L.E., and Miller, J.L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. Journal of the Acoustical Society of America, 92, 723-735

Warren, R.M. (1982). Auditory Perception (NY: Pergamon).

Warren, R. M., & Byrnes, D. L. (1975). Temporal discrimination of recycled tonal sequences: Pattern matching and naming of order by untrained listeners. Perception & Psychophysics, 18, 273-280.

Warren, R.M., and Obusek, C. (1972). Identification of temporal order within auditory sequences. Perception & Psychophysics, 12, 86-90.

Watson, C.S., & Kewley-Port, D. (1988). Some reparks on Pastore (1988). Journal of the Acoustical Society of America, 84, 2266-2270.

Woodworth, R.S., and Schlosberg, (1954). Experimental Psychology. (NY: Holt).

Zera, J., and Green, D.M. (1993). Detecting temporal onset and offset asynchrony in multicomponent complexes. Journal of the Acoustical Society of America, 93, 1038-1052.

Zue, V. W. (1976). Acoustic characteristics of stop consonants: A controlled study (Unpublished PHD Dissertation). Massachusetts Institute of Technology.

Acknowledgements

This research was supported by grants F496209310033 and F49609310327 from the Air Force Office of Scientific Research. The opinions, findings, conclusions, and recommendations are those of the authors and do not necessarily represent those of the granting agency.

Figure Captions

Figure 1. A summary of the hypothetical distributions of response categories in measuring the difference limen for some arbitrary stimulus value along a hypothetical continuum. Panel A illustrates the three expected perceptual categories: a center region (or interval) of uncertainty surrounding perceptual and physical equality, and regions of perceived differences above and below the interval of uncertainty. The center curve is a Gaussian distribution; the other two curves are cummulative Gaussian distributions adjusted so that the sum of ordinates at every point equals a constant. The three category method has responses corresponding to each of the three distributions. Panel B shows the distribution of responses for measuring the DL using the two-category procedure. For any given stimulus value, probability of a given "high" (or "low") response is based upon the probability derived from that distribution in Figure 1a plus 50 percent (for guessing) of the probability from the uncertainty distribution. The 50 percent threshold values from Figure 1a correspond with the 75 percent threshold values in Figure 1b.

Figure 2. The DL for magnitude of order of onset relative to a 5 ms standard onset difference is plotted as a function of the time interval (ISI) between the two stimuli presented on a given trial. The open symbols represent the Rising F2 stimuli and the filled symbols represent the Falling F2 stimuli. The solid lines represent the mean for the two stimulus conditions.

Figure 3. The DL for magnitude of order of onset is plotted as a function of the smaller difference in onset for the stimuli with the Falling F2 component. The two panels separately plot the results obtained under fixed and roving discrimination tasks. The separate symbols represent individual subjects. The two lines are the linear regression solutions for the sets of data for small and large initial differences in onset.

Figure 4. DL results for Rising F2 stimuli (see Figure 3 for details).

Table 1. Summary of Stimulus Parameters for Systematic Evaluation of Cues for Place of Articulation

	Li & Pastore (1992)	Experiment 1	Experiment 2
	Vowel /a/	Vowel /i/	Vowel /a/
Durations (ms)			
Burst	(none)	5	5
Silence	-	5	5
Transition	40	40	50
Steady-State	160	200	250
Vowel Freq (Hz)			
F0	120	120	125
F1	700	379	720
F2	1,220	2,200	1,240
F3	2,600	3,000	2,500
Consonant Freq			
F1-Onset	400	400	200
F2-Onset	600 - 1,800	1,400 - 2,600	1,400 - 2,400
(F2 Step Size)	(200 Hz)	(200 Hz)	(200 Hz)
F3 Onset	1,400 - 3,200	2,400 - 3,400	2,400 - 3,400
(F3 Step Size)	(400 Hz)	(200 Hz)	(200 Hz)
Low Freq Burst	(none)	1.0 - 2.75 kHz	1.0 - 2.75 kHz
High Freq Burst	(none)	2.0 - 4.0 kHz	2.0 - 4.0 kHz

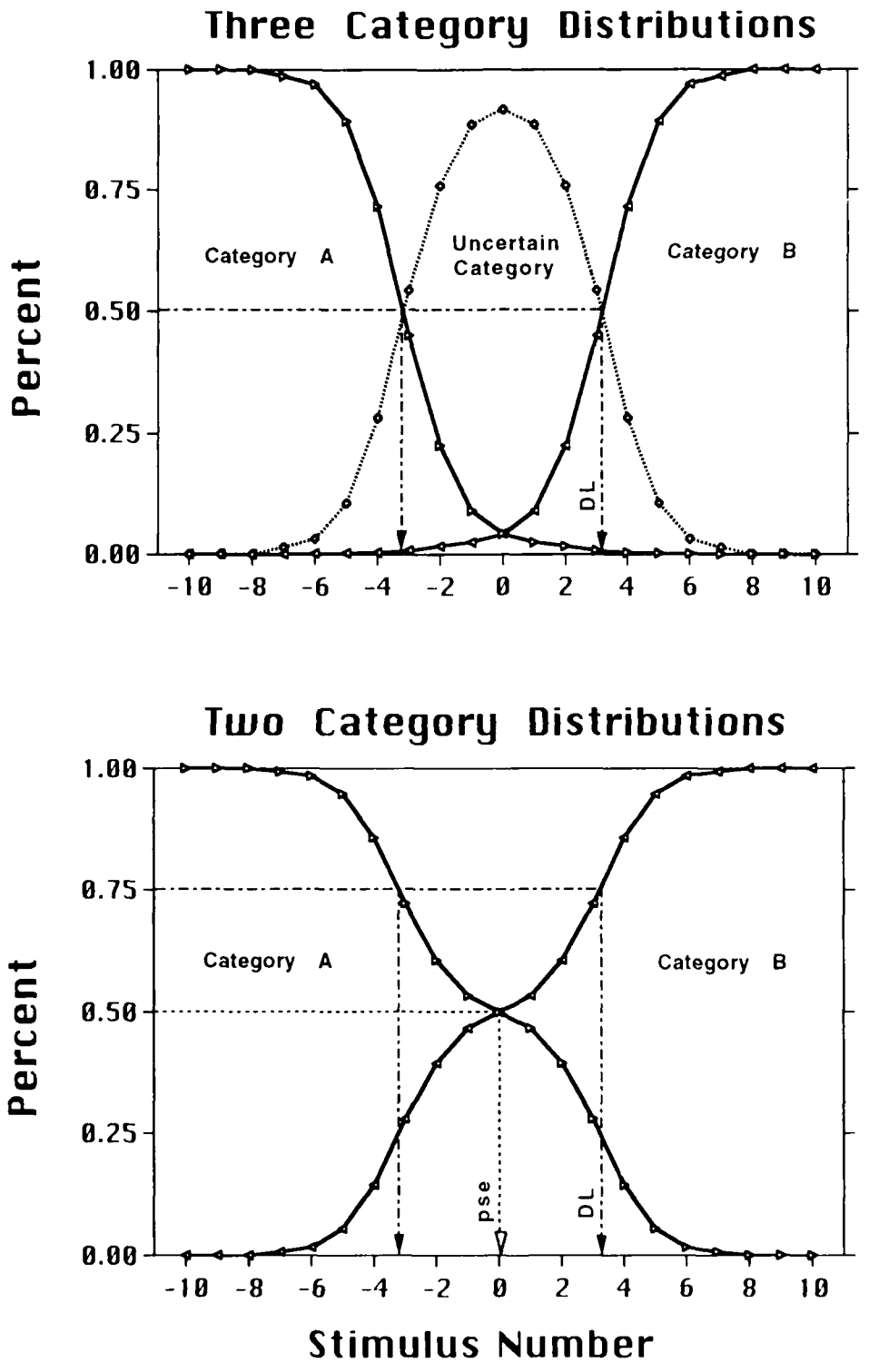

Three Category Distributions
1.00
0.75
0.50
0.25
0.00
Percent
Category A
Uncertain Category
Category B
DL
-10 -8 -6 -4 -2 0 2 4 6 8 10
Two Category Distributions
1.00
0.75
0.50
0.25
0.00
Percent
Category A
Category B
pse
DL
-10 -8 -6 -4 -2 0 2 4 6 8 10
Stimulus Number

Figure 1

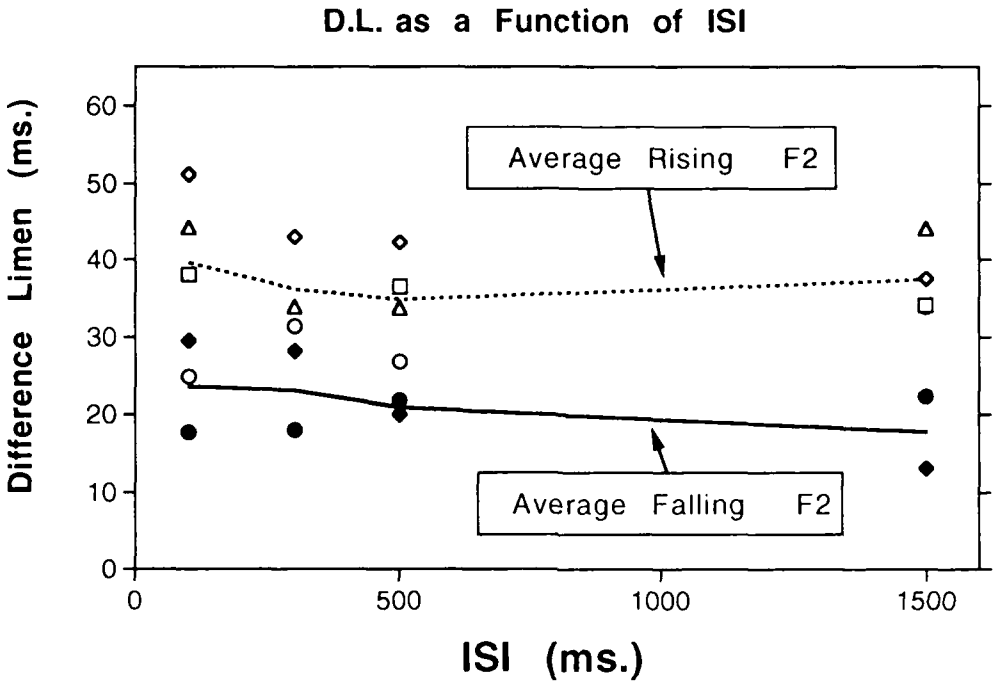

D.L. as a Function of ISI
Difference Limen (ms.)
60
50
40
30
20
10
0
0
500
1000
1500
ISI (ms.)
Average Rising F2
Average Falling F2

Figure 2

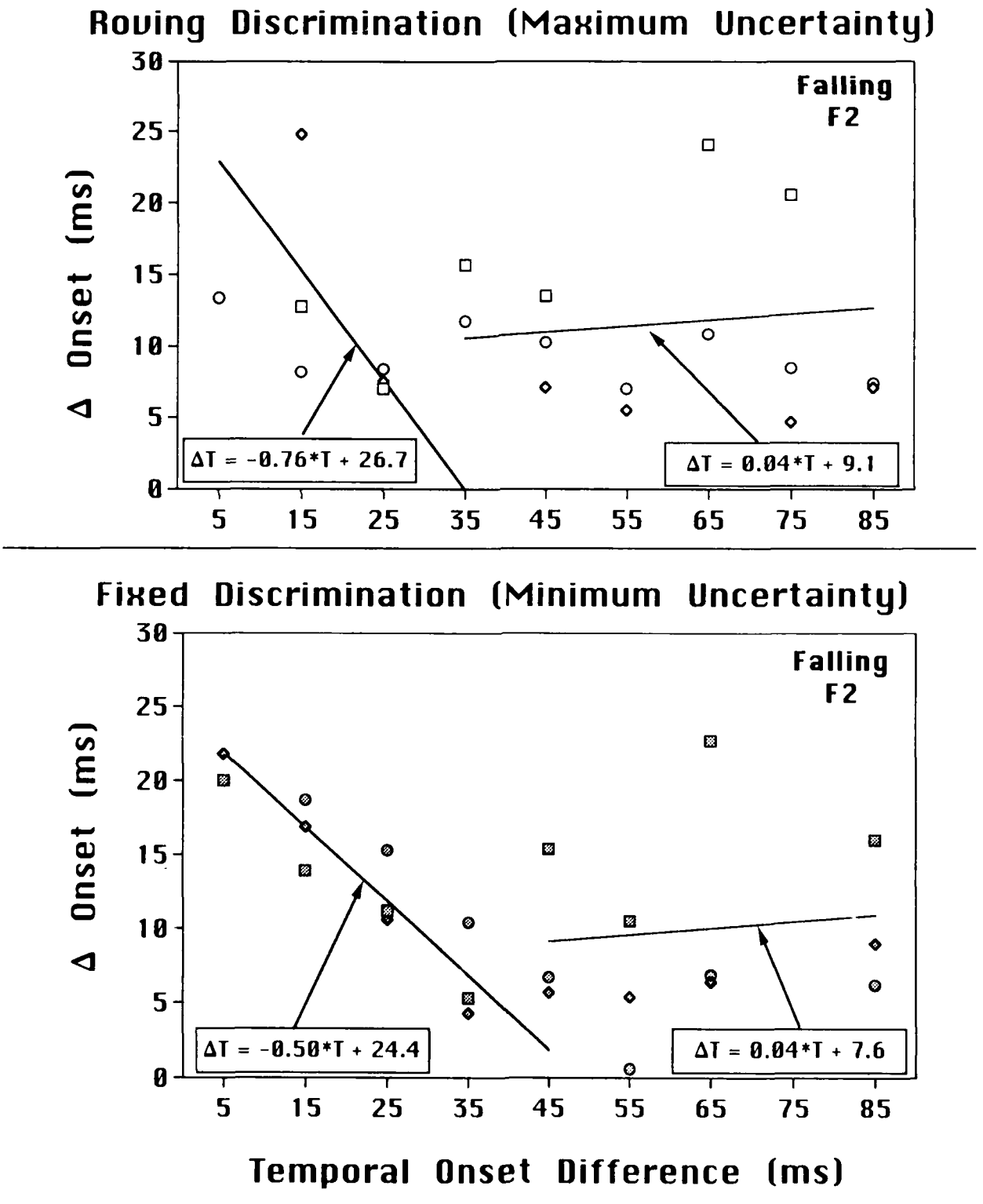
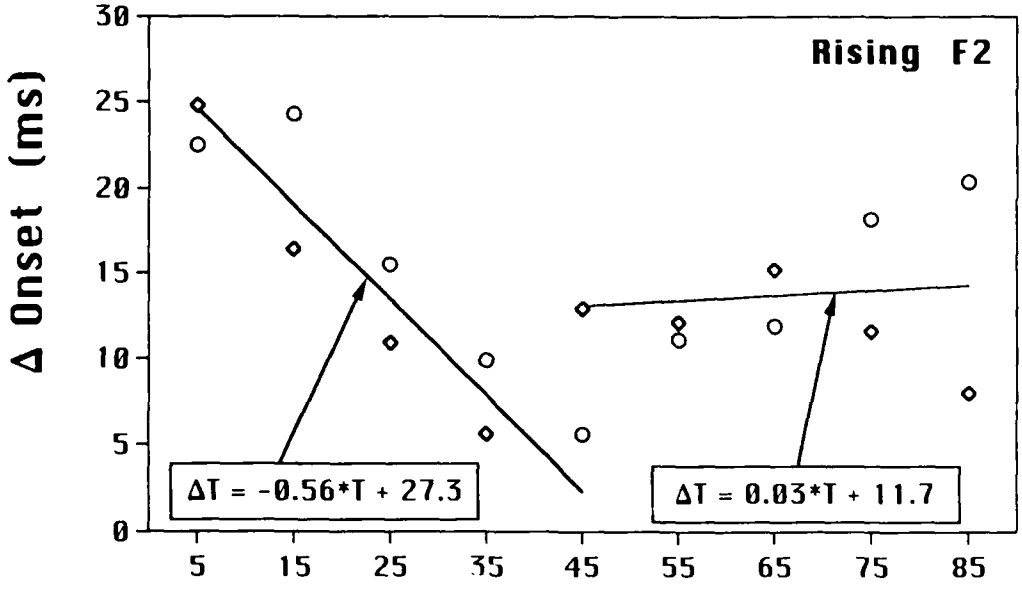

Roving Discrimination (Maximum Uncertainty)
Falling
F2
Δ Onset (ms)
ΔT = -0.76*T + 26.7
ΔT = 0.04*T + 9.1
Fixed Discrimination (Minimum Uncertainty)
Falling
F2
Δ Onset (ms)
ΔT = -0.50*T + 24.4
ΔT = 0.04*T + 7.6
Temporal Onset Difference (ms)

Figure 3

Roving Discrimination (Maximum Uncertainty)



Fixed Discrimination (Minimum Uncertainty)

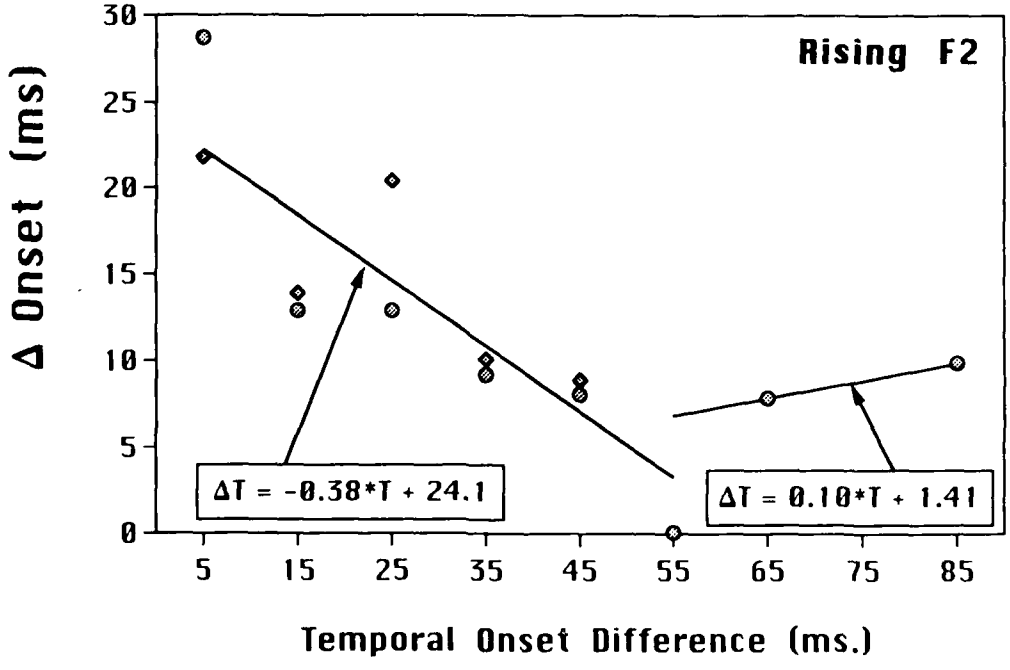


Figure 4

Exploration of the phonetic structure of cues for place of articulation

Richard Pastore, Xiaofeng Li, Jennifer Cho, Barbara Acker, and Shannon Farrington

Abstract

A multi-task, multi-dimensional approach was used to evaluate the nature of perceptual space and the relative importance of auditory cues for the perception of initial position voiced stop consonants of English. For each of several different vowel contexts, stimuli varying systematically in F2- and F3-onset time and the nature of an initial release burst are examined from a number of different perspectives. For each vowel, a within subject design is used, evaluating the stimuli in terms of classification, speeded classification, goodness for each of the possible phoneme categories, and similarity scaling. The scaling results are then submitted to a multi-dimensional scaling analysis. Each of the tasks and stimulus parameters have been the focus of prior investigations of phoneme categories, but have never been combined to provide a complex, systematic picture of perceptual space to provide converging evidence for the nature of cues for phoneme categories.

End of Abstract

Running Heading: Phonetic Structure Exploration

Most research investigating the cues for specific phoneme categories, such as place of articulation, has tended to focus on the location of category boundaries defined along a single physical continuum, largely ignoring the nature, quality, and extent of perception within phonetic categories. More recently, some studies sometimes have been expanded to demonstrate "trading relations," which are really only the simple interaction of a limited range of values for two different physical continua, with the critical dependent variable again being the location of the category boundary defined along one of the two dimensions.

Although this research is relatively simple to conduct, the results of such limited focus research cannot be expected to significantly advance our knowledge about the critical cues which define the perception of speech categories.

Among the more promising alternative approaches to investigating the nature of phonetic perceptual space is the multidimensional scaling of similarity ratings of stimuli (e.g., Pols, van der Kemp, & Plomp, 1969; Carroll & Chang, 1970; Soli, 1982; Bladen & Lindblom, 1981). However, even multidimensional scaling studies have tended to use relatively limited sets of stimulus dimensions. In recent years there also has been a growing interest in the possible role of prototypes or exemplars in defining phonetic categories (e.g., Samuel, 1982; Kuhl, 1991; Volatis & Miller, 1992; Sussman, 1993). These prototype-oriented studies have begun to use several different measures (e.g., goodness rating, discrimination, or selective adaptation) to provide the beginnings of an evaluation of the perceptual structure of stimuli falling within and across phonetic categories. Despite this trend, very few studies have evaluated perception as a function of a number of different stimulus dimensions.

The related studies by Hoffman (1958) and Harris, Hoffman, Delattre, and Cooper (1958) are an excellent example of the evaluation of the perception of place of articulation as a function of several different physical properties, specifically F2-onset frequency, F3-onset frequency, and initial onset bursts. These early studies provide an excellent delineation of the limits (boundaries) of each perceptual category as a function of the three important stimulus variables. Somewhat more recently, Stevens and Bloomstein (1978) evaluated similar variables in the perception of place of articulation, defining the variables more precisely and in terms of more dynamic properties.

The current research uses a number of different types of measures and techniques to provide a more detailed specification of the nature of phoneme perceptual space for categories of initial position voiced stop consonants varying in place of articulation. All of the various measures utilized in this research have been employed separately in past investigations of phoneme perception, but never in combination, and never have been used to provide converging evidence for strong conclusions about phoneme perception. Furthermore, all the basic physical properties of the stimuli also have been subject to extensive investigation, but not at the level or detail of the current research.

The current research evaluates the nature of the perceptual space for the phoneme categories defined by /b/, /d/, and /g/. The stimuli vary in F2- and F3-onset frequency, the presence and nature of an initial burst, and the specification of the following vowel. These variables were identified as important cues in early research by Delattre, et al. (1955), and by Hoffman (1958; Hoffman, et al. 1958). More recent important investigations include those by Stevens and Blumstein (1978), Kewley-Port (1982, 1983), Searey and Shammass (1987), Stevens (1992), and Sussman, McCaffrey, & Matthews (1991). In the current study, the stimuli first are evaluated in terms of speeded classification, thus allowing specification of the location and extent of each phonetic category boundary, as well as obtaining reaction time measures for applying labels. These results also allow for direct comparisons of the current finding with the many published studies focusing on category boundary location along one of the various dimensions investigated. The complete set of stimuli then are evaluated in terms of the relative goodness of the labels "b," "d," and "g." Finally, a subset of stimuli is subjected to the rating of similarity between all possible pairings of stimuli, thus allowing the use of multidimensional scaling techniques to evaluate the underlying dimensionality of perception. The correspondence, or consistency, among the various measures provides an indication of the degree to which the various measures are indicating similar, or different, underlying processes. Furthermore, the nature and shape of the defined perceptual space provide strong indication not only of the critical underlying stimulus dimensions and the symmetry of the perceptual decision mechanism, but also the relavence of prototype versus exemplar (or alternative) models of categorization.

Earlier Research

Li and Pastore (1992) present the initial work in this research project. This study systematically varied the F2- and F3-onset frequency for a set of synthetic CV syllables (without initial burst) based upon the vowel /a/. The stimulus parameter for this study (along with those from Experiment 1 and 2) are summarized in Table 1. The study used an open-ended labeling task, and a speeded classification task on the full set of stimuli plus a goodness rating task, and a similarity rating task on a subset of the stimuli to first provide an overall perspective of the perceptual space for the phonetic categories /b/, /d/, and /g/. This experiment used 14 to 18 subjects per

condition, but different subjects for each condition. The basic finding for this vowel condition was that /b/ is defined by a low F2-onset frequency (F2 < 1100 Hz); thus rising F2 at onset. For higher F2-onset frequencies the perceived category (/d/ or /g/ and almost never "other") depended upon an interaction of F2- and F3-onset frequency. The goodness ratings indicated a difference in the perceptual structure of the three categories, with higher levels of goodness for /d/ being concentrated and for /b/ being relatively uniform and diffuse. A combination of speeded classification and labeling results were used to evaluate predictions from prototype and exemplar models based upon the work of Nosofsky (1991). Both types of models provided excellent predictions, each accounting for 97% of the variance.

Current Research

The research completed to date under the current project represents several improvement over the Li and Pastore study. The stimuli were synthesized using the CSRE 4.0 version of the Klatt synthesizer, with increased variation in F0 and amplitude, especially at offset, to improve the perceived quality of the stimuli. In addition to varying F2- and F3-onset frequency (with finer sampling of F3-onset frequency), three different onset bursts conditions (no burst, low-frequency burst, and high-frequency brust) were added. The stimulus set thus consisted of the factorial combination of values of F2-onset frequency, F3-onset frequency, and three different burst conditions. The two experiments completed to date differed in terms of the vowel, and thus the selection of F2- and F3-onset frequencies. Each experiment used a within-subject design with each of six to eight subjects (normal hearing, native speakers of American English) completing all tasks with the given set of stimuli. The approximately 100 stimuli were subject to speeded classification [open-ended labeling (labels of "b," "d," "g," and "other") with RT measured], separate goodness rating tasks (one each for /b/, /d/, and /g/). Because of the multiplicative growth in number of distinct trials as a function of the number of stimuli (24 stimuli require 576 trials for one presentation of each stimulis pairing; 33 stimuli requires 1089 trials), the evaluation of the similarity between pairs of stimuli was based upon a subset of the original stimuli; this task was repeated five to seven times, each with a new randomization of stimuli. The similarity scaling utilized each of the three burst conditions for a limited set (8 to 11) of different values of F2- and F3-onset frequencies selected on the basis of the goodness rating results. All stimuli were presented binaurally over TDH-49 earphones at a comfortable listening level. Subjects were run singlely or in pairs in commercial sound chambers.

Experiment 1

The first condition was based upon the vowel /i/ and factorially combined seven values of F2-onset frequency with five values of F3-onset frequency and the the three different burst conditions, as summarized in the middle column of Table 1. Figure 1 presents the labeling results in terms of percent correct for the labels "b," "d," and "g" (the label "other" was almost never used) as a function of F2-onset frequency, F3-onset frequency, and type of burst. It is quite clear that /b/ is defined by a combination low F2- and low F3-onset frequencies. Redefining the stimuli in terms consistent with Stevens (1992), /b/ seems to be specified by the upward movement from onset of both the F2 and F3 resonances, but can tolerate a flat F3 resonance when the F2 resonance is clearly rising. Adding a low or high frequency burst decreases the rate of classifying these stimuli as /b/. A burst also shrinks the range of stimuli being acceptable as /b/, removing the falling or flat F3 stimuli from this category. Thus, those stimuli which mixed the direction of F2 and F3 resonances are probably somewhat ambiguous with "b" being the default category label (this conclusion is confirmed by the goodness ratings).

In the absence of a release burst, categorization of the phoneme /g/ seems to be defined in terms of high F2- and F3-onset frequencies, and thus the downward movement from onset of the F2 and F3 resonances. Adding a release burst reduces the /g/ category range to the stimuli with the more sharply falling F2 resonance (F2-onset > 2200). The phoneme /d/ seems to involve F2 and F3 resonances that move in somewhat opposite directions, with this category really existing primarily in the presence of a high frequency noise burst at onset, and even then the category does not seem strong. The pattern of F2- and F3-onset frequencies in defining phonetic categories thus seems to be quite different from that found by Li and Pastore (1991) for the vowel /a/.

Insert Figure 1 and 2 about here

The goodness rating tasks used a scale of 1 (very poor example of the phoneme category) to 7 (excellent example of the phoneme category). The goodness rating results are summarized in Figure 2. It should be noted that there was considerable individual differences in the ratings of the optimum stimulus for any given category, and also in the use of the rating scale to indicate goodness. Such individual differences tended to cause an overall regression of the goodness ratings toward the mean, although the general pattern of results seems relatively consistent across subjects. For /b/ the overall pattern of goodness results is generally equivalent to the labeling results. However, for /g/, the addition of either type initial release burst did not significantly alter goodness of the stimuli with a sharply falling F2 resonance. Moderate levels of goodness for /d/ are found with mixed F2- and F3-resonance changes accompanied by a high frequency burst at onset.

Insert Figure 3 about here

The similarity scaling results for a sampling of 11 values of onset frequencies crossed with the three types of burst conditions (33 stimuli) were subjected to a multi-dimensional scaling analysis which yielded an excellent fit in two dimensions (it is more interesting to obtain a solution which yeilds fewer or different dimensions than the physical variables manipulated. The stimuli in this Figure are coded in terms of burst type (fill of symbols) and relative direction of F2- and F3-onsets (symbol type). It is quite clear from these results that dimension 2 is coding burst type, with the no burst (unfilled symbols) and high burst (darkest filled symbols) stimuli being at the two ends of the dimension and the low burst stimuli (lightly filled symbols) in the center. The two extremes of dimension 1 represent the F2

and F3 resonances both rising (◂) and both falling (▸) together, with the middle portion of the dimension representing a mixture of rising and falling resonances (◆).

Figure 3b replots the multidimensional scaling stimuli in terms of the labeling results. Large, bold symbols designate a very high percentage (90-100%) labeling for the given stimulus, whereas small, mixed stimuli represent approximately equal use of two phonetic labels. Note that dimension 1, coding change in F2- and F3-onset resonances, clearly differentiates /b/ from the other phonetic catergories, and may provide some very small differentiation between the /d/ and /g/; these scaling results are consistent with the classification and goodness results for /b/ in this vowel context. Dimension 2, the nature of the burst, has a small effect on the goodness of the /b/, and tends to provide some differentiation between /d/ and /g/. Therefore, the change in resonant frequency at onset seems to be important for defining /b/ and differentiating it from the other two phonemes, with the nature of the burst primarily distinguishing between /d/ and /g/. However, based upon the levels of labeling and goodness observed, we suspect that other factors not captured in our stimuli are needed to more fully differentiate /d/ from the other voiced phoneme categories.

Experiment 2

The second experiment in this study was still being conducted at the time that this report was prepared. This experiment reexamines the condition with the vowel /a/ studied in the original Li and Pastore experiment, but with a totally new set of synthetic stimuli (with better sampling of F3), and adding the burst conditions utilized in Experiment 1. It is critical for the project that the essential no burst classification results replicate those of Li and Pastore (1992), thus demonstrating the stability of these findings. The addition of the two burst conditions then build upon the basic findings. The current report is based upon the speeded classification and goodness rating results for eight subjects, all of whom currently are being run under the scaling conditions.

Insert Figure 4 and 5 about here

Figure 4 summarizes the labeling results from Experiment 2 for seven of the subjects; one subject produced labeling and goodness rating results which were in places discrepant relative to the other subjects (e.g., assigning goodness rating below 2.0, or labeling percentages below 20%, to stimuli which received rating above 6.0, or percentages above 80%, from all of the other subjects, and visa versa). The no burst classification results for the perception of /b/ are consistent with those reported by Li and Pastore using a completely different sample of stimuli based upon the same vowel. All stimuli with F2-onset frequencies below approximately 1200 Hz are classified as /b/. At higher values of F2-onset frequency the predominant labeling category is /d/, with some shift toward /g/ at low values of F3. The new set of classification results are for the two conditions which add an initial burst. In each of these conditions the presence of an onset burst does not change the labeling of /b/. Furthermore, the rate of labeling for low F2 stimuli appears to be relatively uniform across burst F3-onset frequency. Therefore, the classification results indicate that F2-onset frequency (or direction of change from onset) differentiates /b/ from the other phonetic categories (/d/ and /g/).

At high F2-onset frequencies (falling F2 resonances) adding a low frequency burst has a relatively uniform effect, causing these stimuli to perceive primarily as /g/. Finally, changing to a high frequency burst causes these high F2-onset frequency stimuli to be perceived almost uniformly as /d/. It appears that the F3-onset frequency, at least for the range of values sampled, has very little effect on perception for this vowel. Therefore, the classification results indicate /d/ and /g/ are distinguished by burst-type for both /a/ and /i/.

Figure 5 summarizes the goodness rating results for the seven subjects. Each subject exhibited a median rating of between 6 and 7 for at least one stimulus in each of the three phoneme categories; as well as for a median rating of 1.0, the poorest possible rating, demonstrating that all subjects used the full range of rating values in each rating task. Although stimuli with higher ratings tended to be grouped for individual subjects, the exact location of the highest rating differed somewhat between subjects thus causing the average of the median ratings to be somewhat lower than those typical of individual subjects, but still quite high within categories.

In general, the goodness rating results tend to parallel the labeling results. /b/ is characterized by a low F2-onset frequency and is independent of both an initial burst and the nature of the F3-onset frequency. /g/ and /d/ are characterized by a high F2-onset frequency together with an initial noise burst, with burst type differentiating between the two categories. One difference between the goodness and the classification results is found under the no burst condition for the higher F2 stimuli. Although these stimuli are classified as /d/, none of the stimuli are very good tokens of this category; /d/ seems to be a default labeling category.

The contrast between the current approach and the more traditional approach based upon the location of category boundaries can be illustrated by referring back to the low burst condition in Fig. 4. If one were to hold F3-onset frequency constant at 2000 Hz and vary F2-onset frequency, one would find two discrete categories (/b/ and /g/) with a 50 percent labeling boundary at approximately 1200 Hz. The traditional interpretation of such results would be that F2-onset frequency is a cue which differentiates between /b/ and /g/. As we have seen, a low F2-onset frequency is definitely a cue for /b/ and differentiates it from other phonetic categories, but F2-onset frequency is not an adequate cue for /g/. Expanding upon this example, assume that we now run another labeling condition with F3-onset frequency fixed at 2800 Hz, again varying F2-onset frequency. We here would find a category boundary at approximately 1275 Hz, thus apparently demonstrating a trading relation between F3- and F2-onset frequencies for the contrast between the /b/ and /g/. However, we have already seen that F3-onset frequency has very little effect on phoneme category, or category goodness, for this vowel. Instead, we appear to have relatively simple, straight-forward definitions of categories based upon F2 onset and burst-type. We see these categories when we better capture the complexity of the stimuli (as in Figure 3 and 4), rather than trying to draw strong conclusions based upon small changes in the location of a labeling boundary along a single dimension or slice through the perceptual space.

It is tempting to draw some conclusions about the role of transition for low-frequency formants in defining /b/. However, similar types of results really need to be collected for other vowels before any strong conclusions are conjectured.

Other Experiments

Upon completion of the similarity scaling condition for Experiment 2 we will (1) collect similar data for the /u/ vowel context and re-examine the context of the /i/ vowel using a new set of stimuli to attempt to obtain better exemplars of /d/. At that point we will have sampled vowels from three major front-back locations. At that time we will prepare a major manuscript describing the results obtained to date. In addition, we will explore the perceptual space for these phonetic contrasts (/b/, /d/, and /g/) in the context of other vowels, and for other possible cues for phoneme contrast.

Bibliography

Best, C.T., Morrongiello, B., and Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. Perception and Psychophysics, 29, 191-211.

Bladon, R.A., and Lindblom, B. (1981). Modeling the judgement of vowel quality differences. Journal of the Acoustical Society of America, 69, 1414-1422.

Blumstein, S.E., Isaacs, E., and Mertus, J. (1982). The role of the gross spectral shape as a perceptual cue to place of articulation in initial stop consonants. Journal of the Acoustical Society of America, 72, 43-50.

Blumstein, S.E. and Stevens, K.N. (1979). Acoustic invariance in speech production: Evidence from measurements of the spectral characteristics of stop consonants. Journal of the Acoustical Society of America, 66, 1001-1017.

Carroll, J.D., and Chang, J.J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. Psychometrika, 35, 283-319.

Delattre, P.C., Liberman, A.M., and Cooper, F. S. (1955). Acoustic loci and transitional cues for consonants. Journal of the Acoustical Society of America, 27, 769-773.

Harris, K.S., Hoffman, H.S., Liberman, A.M., Delattre, P.C., and Cooper, F.S. (1958). Effect of third-formant transitions on the perception of the voiced stop consonants. Journal of the Acoustical Society of America, 30, 122-126.

Hoffman, H.S. (1958). Study of some cues in the perception of the voiced stop consonants. Journal of the Acoustical Society of America, 30, 1035-1041.

Kewley-Port, D. (1982). Measurement of formant transitions in naturally produced stop consonant-vowel syllables. Journal of the Acoustical Society of America, 72, 379-389.

Kewley-Port, D. (1983). Time-varying features as correlates of place of articulation in stop consonants. Journal of the Acoustical Society of America, 73, 322-335.

Kuhl, P.K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. Perception and Psychophysics, 50, 93-107.

Lt. X.-F., and Pastore, R.E. (1992). Evaluation of prototypes and exemplars for a phoneme place continuum. In: M.E.H. Schouten (Ed), Audition, Speech and Language. Berlin: Mouton-De Gruyter, 303-308.

Nearey, T.M., and Shammass, S.E. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. Can. Acoust., 15(4), 17-24.

Nosofsky, R.M. (1986). Attention, similarity, and the identification-categorization relationship. Journal of Experimental Psychology: General, 115, 39-57.

Nosofsky, R.M. (1991). Tests of an exemplar model for relating perceptual classification and recognition memory. Journal of Experimental Psychology: Human Perception and Performance, 17, 3-27.

Pols, L.C.W., van der Kemp, L.J. Th., and Plomp, R. (1969). Perceptual and physical space of vowel sounds. Journal of the Acoustical Society of America, 46, 458-467.

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. Psychological Bulletin, 92, 81-110.

Repp, B.H. (1983). Trading Relations Among Acoustic Cues in Speech Perception are Largely a result of Phonetic Categorization. Speech Communication, 2, 341-362.

Samuel, A.G. (1982). Phonetic prototypes. Perception and Psychophysics, 31, 307-314.

Soli, S. (1983). The role of spectral cues in discrimination of voice onset time differences. Journal of the Acoustical Society of America, 73, 2150- 2165.

Stevens, K.N., and Blumstein, S.E. (1978). Invariant cues for place of articulation in stop consonants. Journal of the Acoustical Society of America, 64, 1358-1368.

Stevens, K.N., and Blumstein, S.E. (1981). The search for invariant acoustic correlates of phonetic features. In P.D. Eimas and J.L. Miller (Eds.), Perspectives on the Study of Speech. Hillsdale, NJ: Erlbaum.

Stevens (1992). Colloquium at Cornell University, March 1992.

Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., and Cooper, F.S. (1970). Motor theory of speech perception: A reply to Lane's critical review. Psychological Review, 77, 234-249.

Sussman, H.M., McCaffrey, H.A., and Matthews, S.A. (1991). An investigation of locus equations as a source of relational invariance for stop place categorization. Journal of the Acoustical Society of America, 90, 1309-1325.

Sussman, J.E. (1993). A preliminary test of prototype theory for a [ba]-to-[da] continuum. Journal of the Acoustical Society of America, 93, 2392 (Abstract).

Volaitis, L.E., and Miller, J.L. (1992). Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. Journal of the Acoustical Society of America, 92, 723-735.

Zue, V.W. (1976). Acoustic characteristics of stop consonants: A controlled study. (Unpublished PHD Dissertation). Massachusetts Institute of Technology.

Acknowledgements

This research was supported by grant F496209310033 from the Air Force Office of Scientific Research. The opinions, findings, conclusions, and recommendations are those of the authors and do not necessarily represent those of the granting agency.

Figure Captions

Figure 1. Classification phonemes /b/, /d/ and /g/ in context of vowel /i/ as a function of F3-onset frequency (absyssa of each bar graph), F2-onset frequency (separate rows of bar graphs), and the nature of the initial burst (columns of bar graphs). For each individual stimulus the percent labeling for the categories /b/ (red with light diagonal line), /d/ (light blue with knotted pattern), and /g/ (yellow with brick pattern) are indicated. Because the category "other" was seldom used, the rate of this response is not indicated other than by the sum of the other three labeling rates being less than 100.

Figure 2. Goodness rating results for vowel /i/. The organization of this Figure is equivalent to the classification results in Figure 1. The highest level of goodness is 7.0, and goodness ratings were obtained separately for each of the three phoneme categories.

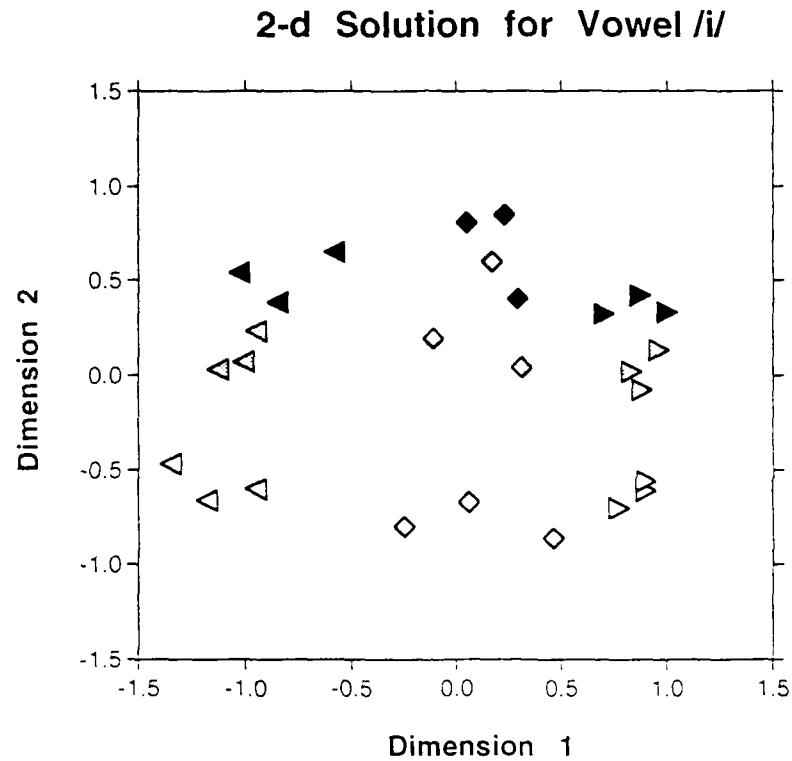
Figure 3. Two-dimensional solution for similarity scaling results is coded in the upper panel in terms of the nature of the F2- and F3-onset resonances and the nature of the burst. The two resonances can be both rising (◂), both falling (▸), or a mixture of rising and falling (◆). The burst can be absent (open symbols), low frequency (light fill), or high frequency (dark fill). The lower panel is coded in terms of the relative frequencies of labeling contained in the classification results. The large symbol indicates a very high rate of labeling for the given category, whereas a small, double symbol indicates between 40 and 60 percent labeling for the two phoneme categories, with the more frequent category listed first.

Figure 4. Classification results for the vowel /a/. (See Figure 1 for description of Figure organization). In this Figure the dark blue category represents the use of the label "other."

Figure 5. Good rating results for vowel /a/ (See Figure 2 for description of organization).

Table 1: Summary of Stimulus Parameters for Systematic Evaluation of Cues for Place of Articulation

	Li & Pastore (1992	Experiment 1	Experiment 2
	Vowel: /a/	Vowel: /i/	Vowel: /a/
Durations (ms.):			
Burst	(none)	5	5
Silence	-	5	5
Transition	40	40	50
Steady-State	160	200	250
Vowel Freq. (Hz):			
F0	120	120	125
F1	700	379	720
F2	1,220	2,200	1,240
F3	2,600	3,000	2,500
Consonant Freq.:			
F1-Onset	400	400	200
F2-Onset	600 - 1,800	1,400 - 2,600	1,400 - 2,400
(F2 Step Size)	(200 Hz.)	(200 Hz.)	(200 Hz.)
F3-Onset	1,400 - 3,200	2,400 - 3,400	2,400 - 3,400
(F3 Step Size)	(400 Hz)	(200 Hz)	(200 Hz)
Low Freq. Burst	(none)	1.0 - 2.75 kHz.	1.0 - 2.75 kHz.
High Freq. Burst	(none)	2.0 - 4.0 kHz.	2.0 - 4.0 kHz


2-d Solution for Vowel /i/
Dimension 2
Dimension 1
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5

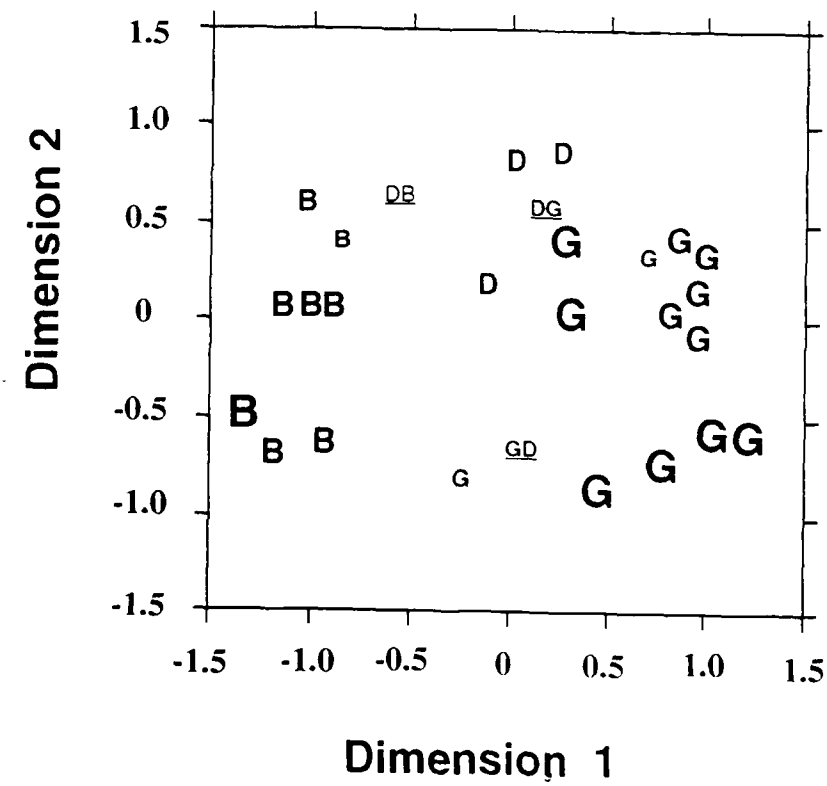
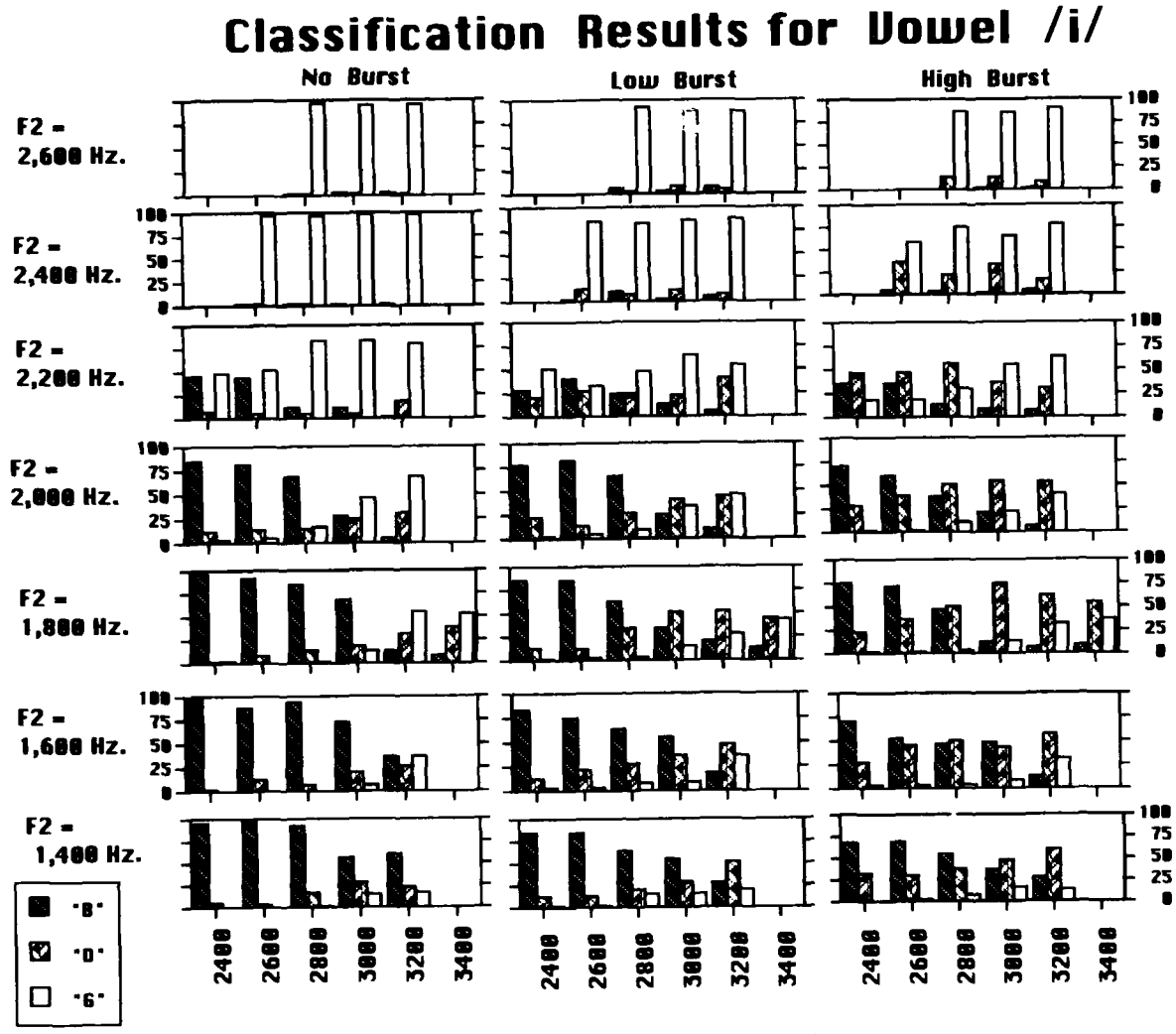
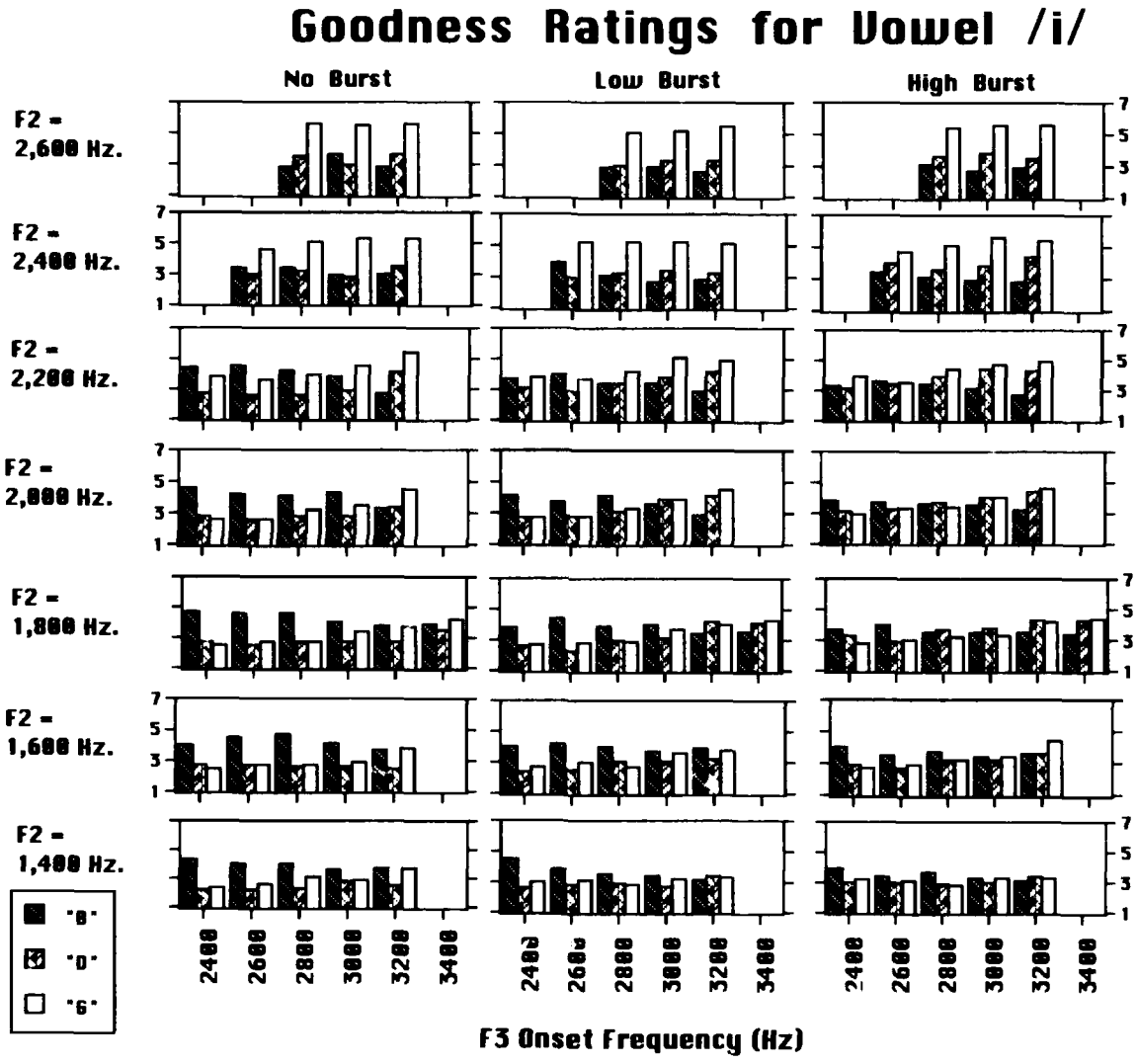
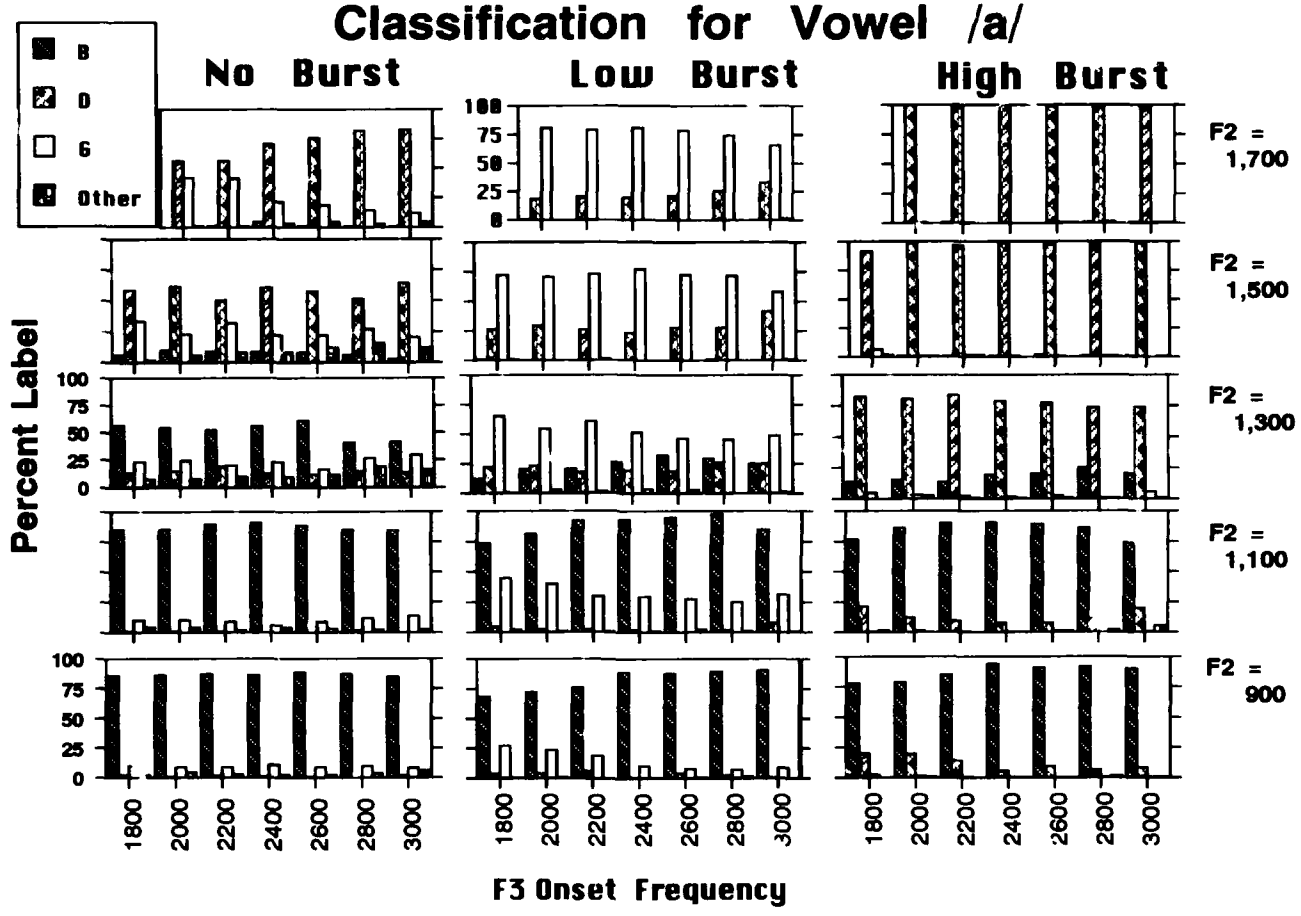

Dimension 2
Dimension 1
1.5
1.0
0.5
0
-0.5
-1.0
-1.5
B
DB
D
DG
G
GD

Figure 3


Classification Results for Vowel /i/
No Burst
Low Burst
High Burst
F2 = 2,600 Hz.
F2 = 2,400 Hz.
F2 = 2,200 Hz.
F2 = 2,000 Hz.
F2 = 1,800 Hz.
F2 = 1,600 Hz.
F2 = 1,400 Hz.
100
75
50
25
0
"B"
"D"
"G"
2400
2600
2800
3000
3200
3400
F3 Onset Frequency (Hz)


Goodness Ratings for Vowel /i/
No Burst
Low Burst
High Burst
F2 = 2,600 Hz.
F2 = 2,400 Hz.
F2 = 2,200 Hz.
F2 = 2,000 Hz.
F2 = 1,800 Hz.
F2 = 1,600 Hz.
F2 = 1,400 Hz.
"B"
"D"
"G"
2400
2600
2800
3000
3200
3400
F3 Onset Frequency (Hz)


Classification for Vowel /a/
No Burst
Low Burst
High Burst
B
D
G
Other
F2 = 1,700
F2 = 1,500
F2 = 1,300
F2 = 1,100
F2 = 900
Percent Label
100
75
50
25
0
1800
2000
2200
2400
2600
2800
3000
F3 Onset Frequency

Goodness Ratings for Vowel /a/